Predictive Modeling of Cholera Outbreaks in Bangladesh

Amanda Koepke

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Volodymyr Minin, Chair

Ira M. Longini, Jr., Chair

Jonathan Wakefield

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

Predictive Modeling of Cholera Outbreaks in Bangladesh

Amanda Koepke

Co-Chairs of the Supervisory Committee:
Associate Professor Volodymyr Minin
Departments of Statistics and Biology

Associate Professor Ira M. Longini, Jr.
Department of Biostatistics

Despite seasonal cholera outbreaks in Bangladesh, little is known about the relationship between environmental conditions and cholera cases. We seek to develop a predictive model for cholera outbreaks in Bangladesh based on environmental predictors. To do this, we estimate the contribution of environmental variables, such as water depth and water temperature, to cholera outbreaks in the context of two different disease transmission models. First, we develop a Bayesian estimation procedure that simultaneously accounts for disease dynamics and environmental variables in a Susceptible-Infected-Recovered-Susceptible (SIRS) model. The entire system is treated as a continuous-time hidden Markov model, where the hidden Markov states are the numbers of people who are susceptible, infected, or recovered at each time point, and the observed states are the numbers of cholera cases reported. We implement a particle Markov chain Monte Carlo algorithm to approximate the posterior distribution of the hidden SIRS model parameters. We test this method using both simulated data and data from Mathbaria, Bangladesh. We use the posterior distribution of the hidden SIRS model parameters to make short-term predictions that capture the formation and decline of epidemic peaks. We demonstrate that our model can successfully predict an increase in the number of infected individuals in the population weeks before the observed number of cholera cases increases, which could allow for early notification of an epidemic and timely allocation of resources. We apply this Bayesian analysis to data from

multiple geographical areas in Bangladesh to test the generalizability of our methods and results. We then expand our analysis of the Mathbaria data to include multiple environmental covariates shifted in time by multiple lags, testing estimation and prediction in the presence of multiple highly correlated predictors. Finally, we add an additional latent water compartment to the hidden SIRS model and explore the difficulties of parameter estimation and cholera outbreak prediction using this complex, but biologically more realistic model for cholera transmission.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Thank you to my advisor Vladimir Minin, for everything you taught me and for all of the time you spent supporting me in my research. I never realized that being a great advisor requires such a delicate balance between acting as both a driving and an encouraging force until I saw how well you executed this balance. I would like to thank Ira Longini and Betz Halloran for involving me in such an interesting project and supporting me through my graduate school career. Your instruction and support have been invaluable. My appreciation goes to Ira for his guidance over the years and for giving me the opportunity to travel to Bangladesh and meet our collaborators. My experiences there will forever have an impact on me. My continued thanks go to Betz for her mentoring and excitement about my research. Thanks also go to Jon Wakefield, for bringing a different statistical perspective and a healthy dose of humor to my committee. Thank you to everyone in Vladimir's research group, from whom I learned so much about stochastic processes and genetics. Thank you also to CSQUID, whose members taught me so much about modeling infectious disease data. I gratefully acknowledge my funding from the NIH grants R01-AI039129 and U01-GM070749, and collaborators at the ICDDR,B who collected and processed the data. Thank you to the classmates, officemates, and friends that have been there for me on this journey. I could not have accomplished everything I have without the support, guidance, and love of my family. And to my husband Lucas, thank you for being a constant source of support and confidence.

DEDICATION

To my parents

Chapter 1

OVERVIEW OF CONTRIBUTIONS

1.1 Cholera

Cholera is an acute diarrhoeal infection which causes severe dehydration and can be fatal if left untreated; it is caused by ingestion of the bacterium *Vibrio cholerae* [Sack et al., 2004, Harris et al., 2012]. According to the World Health Organization, there are an estimated 3-5 million cases and 100,000-120,000 cholera related deaths annually [WHO, 2014]. It is endemic in many countries and areas that lack access to adequate sanitation, such as Bangladesh. In Bangladesh alone there are an estimated 352,000 cases and 3,500 to 7,000 deaths annually [International Vaccine Institute, 2012].

In Bangladesh, outbreaks occur seasonally [Swaroop and Pollitzer, 1955, Glass et al., 1982, Sack et al., 2003, Huq et al., 2005, Koelle and Pascual, 2004, Koelle et al., 2005, Longini et al., 2002]. Since *Vibrio cholerae* can be detected in the environment year round [Huq et al., 1990, Colwell and Huq, 1994], it is hypothesized that environmental forces contribute to the spread of cholera. Thus, environmental covariates could be used to predict an upcoming outbreak. However, little is known about what environmental covariates are important for cholera outbreak prediction. In the past, analyses have found that different covariates are significant in different geographical areas and at different lags, leaving no consistent model for the prediction of cholera epidemics [Huq et al., 2005].

We seek to understand the dynamics of cholera and to develop a model that will be able to predict outbreaks several weeks in advance. If the timing and size of a seasonal epidemic could be predicted reliably, vaccines and other resources could be allocated effectively to curb the impact of the disease. A mobile vaccine stockpile could prevent many infections, but implementing effective vaccination strategies requires an understanding of the disease

dynamics.

1.2 Challenges of modeling seasonal cholera outbreaks

We want to incorporate realistic, non-linear dynamics into a transmission model for cholera. Statistical inference in these models is very difficult, as the likelihood may be intractable.

Many approximate methods exist to overcome likelihood intractability, but these methods often make unrealistic assumptions about the transmission process or the available data. Discrete time methods, such as the auto-Poisson model of Held et al. [2005] and the time-series Susceptible-Infected-Recovered (SIR) model of Finkenstädt and Grenfell [2000], require evenly spaced data. Disease transmission occurs in continuous time, so it is more realistic to have a model that can deal with unevenly spaced observations. The continuous-time approach of Cauchemez and Ferguson [2008] necessitates the assumption that the transmission parameter and number of susceptibles remain relatively constant within an observation period, which seems unrealistic for modeling endemic cholera with seasonal outbreaks. We hypothesize that disease dynamics, such as the fluctuation in the number of susceptible individuals over time, play a critical role in the overall transmission process.

To avoid these approximations, both maximum likelihood and Bayesian methods have been developed to allow inference under non-linear transmission models. Maximum likelihood based statistical inference techniques have been developed which use Monte Carlo to allow maximization of the likelihood without explicitly evaluating it [He et al., 2010, Bretó et al., 2009, Ionides et al., 2006, Bhadra et al., 2011]. To implement a Bayesian approach, the particle filter Markov chain Monte Carlo (MCMC) methods developed by Andrieu et al. [2010] require only an unbiased estimate of the likelihood. This approach has recently been used to study infectious disease time-series by Dukic et al. [2012] and Rasmussen et al. [2011], and it is the framework we use in this thesis.

1.3 Contribution of this thesis

We develop two different compartmental models for cholera transmission: a hidden Susceptible-Infected-Recovered-Susceptible (SIRS) model and a hidden Susceptible-Infected-Water-Recovered (SIWR) model. Environmental covariates, which are hypothesized to affect the proliferation

and decline of *V. cholerae* in the water supply, are incorporated into both models. Both models also include two possible avenues for transmission: direct contact with infected individuals or contact with contaminated water.

In the hidden SIRS model, possible mechanisms for infectious contact between infected individuals and susceptible individuals include both direct person-to-person transmission of cholera and consumption of water that has been contaminated by infected individuals. A separate environmental force of infection, a function of the environmental covariates, also acts directly on susceptibles and represents the force of infection from the natural growth and decline of *V. cholerae* in the water.

In contrast, the SIWR model includes an environmental reservoir effect. In this model, the only mechanism for infectious contact is direct person-to-person transmission of cholera. Infected individuals excrete *V. cholerae* directly into the environmental reservoir, and the environmental covariates are incorporated into the growth of this reservoir as well. The force of infection from the contaminated water then acts separately on susceptibles. This environmental reservoir effect is more biologically accurate for the disease, but creates difficulties in estimation. We explore these difficulties.

Our hidden SIRS and SIWR models borrow features from previous cholera models. The models of Codeço [2001] and Hartley et al. [2005] include the environmental reservoir but exclude the possibility for direct person-to-person transmission of cholera. Tien and Earn [2010] and Eisenberg et al. [2013b] develop an SIWR very similar to ours, with both direct person-to-person transmission and an environmental route of transmission, and explore parameter identifiability, but they do not include environmental covariates in their model. Eisenberg et al. [2013a] extend this framework to include a rainfall data forcing function in the rate of transmission from the environment. Our model instead makes the rate of growth for the concentration of *V. cholerae* in the water be a function of environmental covariates. We believe that our SIWR model accurately reflects current understanding of cholera transmission dynamics.

We develop a biologically realistic model of cholera transmission. Our continuous-time framework allows easy incorporation of data with irregular observation times and for greater parameter interpretability and comparability to models based on deterministic differential

equations. We incorporate multiple environmental covariates that are local to the area where cholera cases are observed, rather than looking at large scale covariates such as rainfall and the El Niño Southern Oscillation [Koelle and Pascual, 2004, Koelle et al., 2005, Eisenberg et al., 2013a]. We develop a framework for covariate selection in order to determine which covariates are most important for prediction. We also compare information from different geographical areas in Bangladesh, looking for similar combinations of predictive covariates.

In this thesis, we use a particle MCMC method, as described by Andrieu et al. [2010], to sample from the posterior distribution of the parameters of our hidden Markov models given the observed data. Our particle marginal Metropolis-Hastings (PMMH) algorithm for hidden SIRS and SIWR models is available as an R package at `https://github.com/vnminin/bayessir`.

1.4 Application

We use the above methodology to explore the relationship between several possible environmental predictors and cholera outbreaks in several thanas (administrative subdistricts with a police station) in Bangladesh. Assuming a hidden SIRS model, estimation and prediction are first tested with simulated data. Using data on cholera incidence and environmental data collected from Bangladesh, we first examine two covariates, water depth and water temperature, in one thana that has been observed for six years, Mathbaria. We then add more covariates to the Mathbaria analysis to see what problems this creates in estimation and prediction. We then extend the two covariate analysis to other thanas with only three years of data. We identify important covariates for prediction, and explore which predictive covariates are consistent across thanas. Finally, we use simulated data and the Mathbaria data to estimate the parameters of the SIWR model using our Bayesian framework.

Chapter 2

INTRODUCTION

2.1 Infectious disease modeling

Compartmental models of disease transmission model the spread of a disease in a population [Kermack and McKendrick, 1927]. These models divide individuals into compartments based on disease status [May and Anderson, 1991, Keeling and Rohani, 2008]. A population is comprised of individuals susceptible to the disease (Susceptible), those who have been infected but are not yet spreading the disease (Exposed), those who are infectious (Infectious), and those who have recovered and are immune to further infection (Recovered or Removed). This leads to different types of epidemic models, such as the Susceptible-Exposed-Infectious-Recovered (SEIR) model. A Susceptible-Infectious-Recovered (SIR) model assumes that the disease has no latent period, so everyone who is infected is infectious. If there is the possibility for either no immunity to be conferred after infection or loss of immunity, a Susceptible-Infectious-Susceptible (SIS) or Susceptible-Infectious-Recovered-Susceptible (SIRS) model is used. The type of compartmental model is chosen based on the characteristics of the disease.

Individuals move between compartments at different rates. Consider a simple SIR model where susceptible individuals are exposed to infectious individuals and become infectious at rate β, and infectious individuals recover at rate γ. In other words, β is the infectious contact rate and γ is the recovery rate. The following set of differential equations describe how the number of individuals in the three compartments change over time:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= -\beta SI, \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta SI - \gamma I, \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I. \end{aligned}$$

Here S is the number of susceptible individuals, I is the number of infectious individuals, R is the number of recovered individuals, and the population size is $N = S + I + R$. The basic reproductive number, $R_0 = (\beta \times N)/\gamma$, is the average number of secondary cases caused by a typical infected individual in a population that consists of both susceptible and immune individuals and is useful in determining the epidemic potential of a disease [Diekmann et al., 1990].

Differential equations can also be used to describe the behavior of a more complicated model, such as the SIWR model for cholera transmission of Tien and Earn [2010]. This model includes a water compartment (W) which quantifies the concentration of *Vibrio cholerae* in the water. Infected individuals excrete *V. cholerae* into the environment, so susceptible individuals are infected through both contact with the environment and direct contact with infected individuals. This is expressed through the following modified set of differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= \mu N - b_W SW - b_I SI - \mu S, \\
\frac{\mathrm{d}I}{\mathrm{d}t} &= b_W SW + b_I SI - \gamma I - \mu I, \\
\frac{\mathrm{d}W}{\mathrm{d}t} &= \alpha I - \xi W, \\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R,
\end{aligned}$$

where b_I is the infectious contact rate, b_W is the rate at which susceptible individuals become infected due to exposure with contaminated water, and γ is the recovery rate. The rate of birth and non-disease related death in this population are both assumed to be equal to μ. Infected individuals excrete cholera into the environment at rate α and pathogen concentration decays at rate ξ. Similar modifications can be made to describe the particular dynamics of any disease.

Using differential equations is a classical approach to modeling disease dynamics over time. However, these deterministic systems lack the random nature of realistic population events, especially in small populations. Sometimes diseases randomly die out, not reaching their deterministic equilibrium. A more realistic way to study diseases dynamics [Garnett

et al., 2011] in a population is through stochastic methods, using stochastic models that retain some properties of their deterministic counterparts, but include the more realistic stochastic nature of the disease dynamics. To do this, we use continuous time Markov chains (CTMCs).

2.2 Stochastic processes and simulation

A stochastic process is a sequence of random variables that evolve, for example, over time. A Markov process is a stochastic process that satisfies the Markov property, which states that all future behavior of the process depends only on the current state of the system. Markov processes can evolve over discrete or continuous state spaces, in discrete or continuous time. Compartmental models evolve continuously through time over a discrete state space, so we focus on continuous time, discrete space Markov chains. Continuous time, continuous state models of disease dynamics exist, such as the stochastic differential equation-based methods of Ionides et al. [2006] and Bhadra et al. [2011]; we do not go into the details of these models.

Consider times t_i for observations $i \in \{0, 1, ..., n\}$, where $0 \leq t_0 < t_1 < \ldots < t_n$, and discrete states $x_0, \ldots, x_n, y$. Then, for $s \geq 0$, a continuous time, discrete space Markov chain has the property that

$$P(X_{t_n+s} = y | X_{t_n} = x_n, X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_0} = x_0) = P(X_{t_n+s} = y | X_{t_n} = x_n).$$

This implies that, given the present state of the system, the past and the future states are independent. A Markov chain is homogeneous if, for all $s \geq 0$,

$$P(X_{t_n+s} = y | X_{t_n} = x) = P(X_{t_0+s} = y | X_{t_0} = x).$$

Let $P(X_{t_n+s} = y | X_{t_n} = x) = P(x, y, s)$ be the probability of moving from state x to state y from time t_n to time $t_n + s$. For a discrete state space $x_0, \ldots, x_n \in E$, we can write the probabilities of transitioning between all possible states as a matrix

$$\mathbf{P}(s) = \begin{pmatrix} P(x_0, x_0, s) & \dots & P(x_0, x_n, s) \\ \vdots & \ddots & \vdots \\ P(x_n, x_0, s) & \dots & P(x_n, x_n, s) \end{pmatrix}.$$

The transition probability matrix $\mathbf{P}(s)$ is stochastic, meaning that its elements are nonnegative and its rows sum to one. A distribution $\boldsymbol{\pi}$ is a stationary distribution of the Markov chain governed by $\mathbf{P}(s)$ if $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T \mathbf{P}(s)$. The Kolmogorov forward equations [Kolmogorov, 1931] are defined as

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \mathbf{P}(t)\mathbf{Q} \tag{2.1}$$

where $\mathbf{Q} = \{q_{x,y}\}_{x,y \in E}$ is the infinitesimal generator and consists of the transition rates

$$q_{x,x} = \lim_{s \to 0^+} \frac{1 - P(x, x, s)}{s}$$

and

$$q_{x,y} = \lim_{s \to 0^+} \frac{P(x, y, s)}{s},$$

if $q_{x,x} < \infty$ and $q_{x,x} = \sum_{y \neq x} q_{x,y}$ $\forall x \in E$. With the initial condition $\mathbf{P}(0) = \mathbf{I}$, the full dynamics of the process are the solution to the differential equation (2.1), $\mathbf{P}(t) = \exp(\mathbf{Q}t)$.

For large state spaces, exponentiating the matrix $\mathbf{Q}$ can be computationally prohibitive. Keeling and Ross [2008] demonstrate the difficulties of calculating these transition probabilities for compartmental models of disease transmission. They implement an exact stochastic continuous-time, discrete-state method for evolving Markov processes, using the deterministic Kolmogorov forward equations to express the probabilities for being in all possible states. Solving this very large set of differential equations describes the entire behavior of the system and works well for small population sizes ($N < 1000$) and SIS or SIR dynamics. However, for large populations or more complex compartmental models this approach is computationally infeasible; for even a simple SIR model, the size of the state space grows on the order of N^2. In addition, if the rates are not constant across time, solving the above

system of differential equations becomes even more challenging. Another way to study stochastic disease dynamics using CTMCs would be to simulate many possible realizations of the process.

2.2.1 Simulating homogeneous SIRS

Much work has been done on simulating stochastic processes, CTMCs in particular, in the area of chemical kinetics. Gillespie developed two methods for exact stochastic simulation of chemical process trajectories with constant rates: the direct method [Gillespie, 1977] and the first reaction method [Gillespie, 1976]. The direct method simulates the time to the next event and then determines which event happens at that time. The first reaction method calculates the time to the next reaction for each of the possible events, and the minimum time to next reaction determines the next step of the chain.

Consider a modified version of the SIR model described in section 2.1:

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= \mu R - (\beta I + \alpha)S, \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= (\beta I + \alpha)S - \gamma I, \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \mu R, \end{aligned}$$

where β is the infectious contact rate, α is an additional force of infection, γ is the recovery rate, and μ is the rate at which immunity is lost. Using the direct method, we can think of our CTMC as a chemical system with different reactions corresponding to transitions between disease compartments in this SIRS model. We model the number of susceptible, infectious, and recovered individuals at time t, $\mathbf{X}_t = (S_t, I_t, R_t)$, as a Markov process [Taylor and Karlin, 1998] with infinitesimal rates

$$\lambda_{(S,I,R),(S',I',R')} = \begin{cases} (\beta I + \alpha)S & \text{if } S' = S - 1,\ I' = I + 1,\ R' = R, \\ \gamma I & \text{if } S' = S,\ I' = I - 1,\ R' = R + 1, \\ \mu R & \text{if } S' = S + 1,\ I' = I,\ R' = R - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the three reactions have the rate functions $h_1(\boldsymbol{X}_t) = (\beta I_t + \alpha)S_t$, $h_2(\boldsymbol{X}_t) = \gamma I_t$, and $h_3(\boldsymbol{X}_t) = \mu R_t$, corresponding to the infinitesimal rates of the CTMC. Then the time to the next reaction, τ, has an exponential distribution with rate $\lambda = h_1(\boldsymbol{X}_t) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$, and the kth reaction occurs with probability $h_k(\boldsymbol{X}_t)/\lambda$, for $k = \{1, 2, 3\}$.

The first reaction method instead simulates the time τ_k that the kth reaction happens for $k = \{1, 2, 3\}$, given no other reactions happen in that time. Then the time to the next reaction $\tau = \min_k(\tau_k)$, and the reaction with the reaction time equal to τ is the event that happens.

Both the direct method and the first reaction method work only for homogeneous Markov chains. If we want to assume that the additional force of infection, α, varies over time, the associated Markov chain is inhomogeneous and we must account for the fact that the transition rate could change before the next reaction occurs.

2.2.2 Simulating inhomogeneous SIRS

Gibson and Bruck [2000] introduce the next reaction method, an efficient exact algorithm to simulate stochastic chemical systems. They extend this next reaction method to include time-dependent rates and non-Markov processes. Anderson [2007] deviates from these methods a bit, using Poisson processes to represent the reaction times, with time to next reaction given by integrated rate functions. This leads to a more efficient modified next reaction method which they extend to systems with more complicated reaction dynamics.

Using the methods described by Gibson and Bruck [2000] and Anderson [2007], to incorporate a time-varying force of infection into the SIRS model we must integrate over the rate function $h_1(\boldsymbol{X}_t, s) = (\beta I_t + \alpha(s))\, S_t$. Thus, to find the time τ_1 that the first reaction happens, given no other reactions happen in that time, we generate $u \sim \text{Uniform}(0, 1)$ and solve

$$\int_t^{\tau_1} h_1(\boldsymbol{X}_t, s)ds = \ln(1/u)$$

for τ_1. Since the other two reactions have no time-varying parameters, we can solve for τ_2 and τ_3, the reaction times of the second and third reactions, using the methods of the

previous section. Then we can continue, using the first reaction method to simulate the process.

We simplify this approach by assuming that the time-varying force of infection, $\alpha(t)$, remains constant each day. We define daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0 + 1, \ldots, t_n - 1\}$, and $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Then we can take advantage of the memoryless property of exponentials and propagate the chain forward in daily increments. Thus, we use the direct method, but when the time to next event exceeds the right end point of the current interval A_i, we restart CTMC simulation from the beginning of the interval A_{i+1} using $\alpha_{A_{i+1}}$ in the waiting time distribution rate $\lambda(\alpha_A) = h_1(\boldsymbol{X}_t, \alpha_A) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$, so $\tau \sim \text{Exp}(\lambda(\alpha_A))$. This modified Gillespie algorithm is depicted and detailed in Figure 2.1.

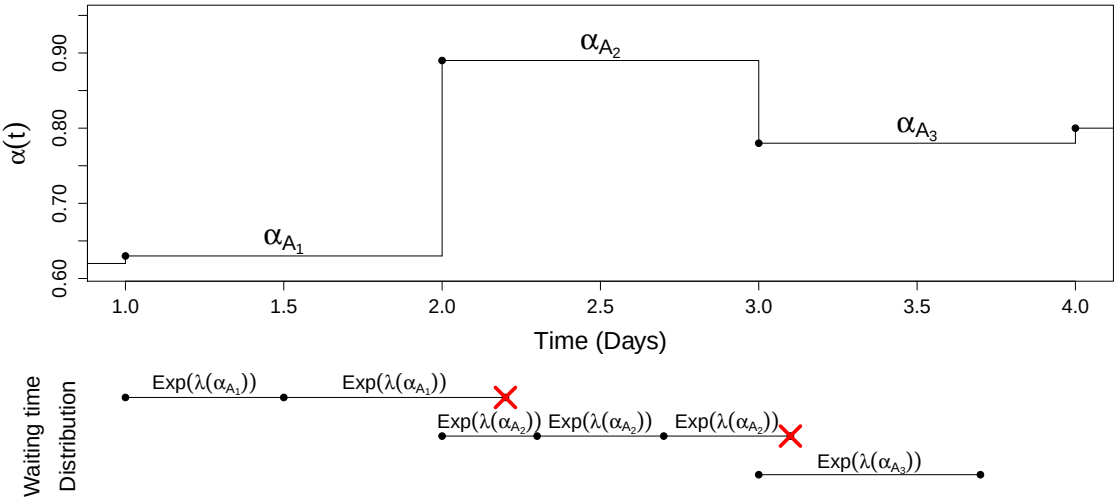

Figure 2.1: Depiction of the modified Gillespie algorithm. We assume the additional force of infection, $\alpha(t)$, is a step function which changes daily. Daily time intervals are denoted by $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n-1\}$, so $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Starting at time $t = 1$, the waiting time to the next event, τ, has an exponential distribution with rate $\lambda(\alpha_{A_1}) = h_1(\boldsymbol{X}_t, \alpha_{A_1}) + h_2(\boldsymbol{X}_t) + h_3(\boldsymbol{X}_t)$. In the depiction, $\tau = 0.5$. The simulated waiting time plus the current time, $t^* = t + \tau$, remains in the interval A_1, so we use t^* as the next time in our CTMC and propagate $\boldsymbol{X}_t$ forward at that time using Gillespie's direct method. Since we are still in the interval A_1, we again simulate the time to the next event as an exponential random variable with rate $\lambda(\alpha_{A_1}) = h_1(\boldsymbol{X}_{t^*}, \alpha_{A_1}) + h_2(\boldsymbol{X}_{t^*}) + h_3(\boldsymbol{X}_{t^*})$. In this iteration, the waiting time plus the current time, $t^* + \tau$, exceeds the boundary of the interval A_1, so we discard this simulated waiting time τ. Using the memoryless property of exponentials, we restart our simulation from the beginning of the interval A_2 using the new $\alpha(t)$ value, α_{A_2}. We continue in this manner until we have simulated the Markov process $\boldsymbol{X}_t$ up to time t_n.

2.2.3 Tau-leaping

The exact algorithms work for small populations, but for large state spaces these methods require a prohibitively long computing time. This is a common problem in chemical kinetics literature, where an approximate method called the tau-leaping algorithm originated [Gillespie, 2001, Cao et al., 2005]. This method simulates CTMCs by jumping over a small amount of time τ and approximating the number of events that happen in this time using a series of Poisson distributions. As τ approaches zero, this approximation theoretically approaches the exact algorithm. The value of τ must be chosen such that the rates remain roughly constant over the period of time; this is referred to as the "leap condition".

Specifically, for our simulation, using the methods outlined in Cao et al. [2005], we define the rate functions $h_1(\boldsymbol{X}_t) = (\beta I_t + \alpha(t))\, S_t$, $h_2(\boldsymbol{X}_t) = \gamma I_t$, and $h_3(\boldsymbol{X}_t) = \mu R_t$, corresponding to the infinitesimal rates of the CTMC. Then $k_1 \sim \text{Poisson}(h_1(\boldsymbol{X}_t)\tau)$ represents the number of infections in time $[t, t+\tau)$, $k_2 \sim \text{Poisson}(h_2(\boldsymbol{X}_t)\tau)$ represents the number of recoveries in time $[t, t+\tau)$, and $k_3 \sim \text{Poisson}(h_3(\boldsymbol{X}_t)\tau)$ represents the number of people that become susceptible to infection in time $[t, t+\tau)$. We make the assumption that the time-varying force of infection, $\alpha(t)$, remains constant each day. We define daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, and $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Using $\tau = 1$ day, our rates now remain constant within each tau jump.

2.2.4 Selecting Tau

Unchecked, tau-leaping can lead to negative population sizes in a compartment if the compartment has a low number of individuals. To avoid this, we use a simplified version of the modified tau-leaping algorithm presented by Cao et al. [2005]. If the population of a compartment is lower than some prespecified critical size, a single step algorithm (like the Gillespie algorithm) is used until the population gets above that critical size. If the size of the compartment is not critically low but the current value of τ still produces a negative population, we reject that simulation and try again with a smaller τ (reduced by a factor of 1/2). The subsequent value of τ is picked based on how long the current daily time-varying force of infection remains constant. We choose a value of τ that simulates what happens

during the remainder of the day, until the value of the transition rate changes. This modified tau-leaping algorithm is depicted and detailed in Figure 2.2.

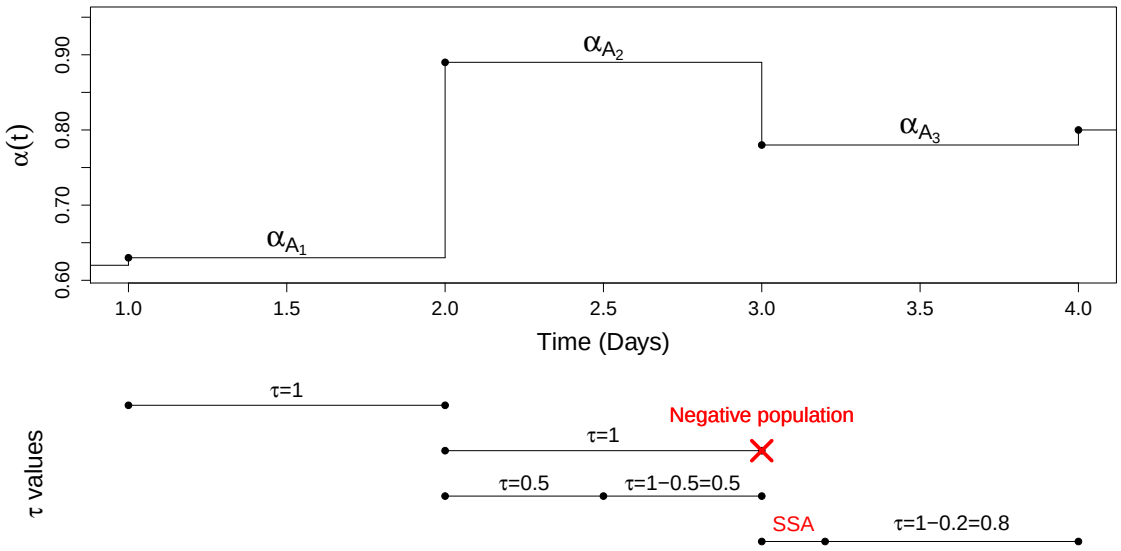

Figure 2.2: Depiction of the modified tau-leaping algorithm. We assume the additional force of infection, $\alpha(t)$, is a step function which changes daily. Daily time intervals are denoted by $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, so $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. As a default, we use $\tau = 1$ day. Starting at time $t = 1$, we simulate the changes in compartment populations over the interval $t \in [1, 2)$. At time $t = 2$, we again use $\tau = 1$ day to simulate the changes over the interval $t \in [2, 3)$. This value of τ produces a negative population so we reject that simulation and try again with a smaller τ (reduced by a factor of $1/2$). The next value of τ is then calculated based on how long the current daily time-varying force of infection remains constant, so $\tau = 0.5$. At time $t = 3$, the population of a compartment is lower than some prespecified critical size, so a single step algorithm (SSA), in our case the Gillespie algorithm, is used until the population gets above that critical size. Once the compartment populations are all above the critical size again, at time $t = 3.2$, the subsequent value of τ is again picked based on how long the current daily time-varying force of infection remains constant, so $\tau = 0.8$.

2.2.5 Binomial tau-leaping

Another solution to the negative population size problem is to use Binomial tau-leaping [Chatterjee et al., 2005, Tian and Burrage, 2004], which further approximates k_j as a bi-

nomial random variable with mean $h_j(\boldsymbol{X}_t)\tau$ and upper limit chosen such that k_j cannot be large enough to simulate a negative population. We opt instead to use the simplified version of the modified tau-leaping algorithm.

2.3 Hidden Markov models

While the underlying dynamics of the disease are described by $\mathbf{X}_t$, we do not directly observe these quantities. Typically, only a random fraction of the number of infectious individuals at each time point, y_t, is observed. This fraction depends on the number of symptomatic infected individuals that seek treatment and get reported (the reporting rate). Thus, y_t is the number of observed infections at time $t \in \{t_0, t_1, ..., t_n\}$. Given $\boldsymbol{X}_{t_i}$, y_{t_i} is independent of the other observations and other hidden states. This hidden Markov model is depicted in Figure 2.3.

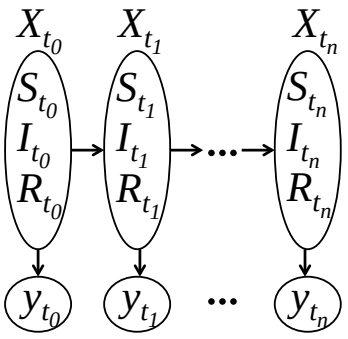

Figure 2.3: Hidden Markov model (HMM) for SIRS. At each observed time point $t \in \{t_0, t_1, ..., t_n\}$, we observe only y_t, which depends on the unobserved state vector $\mathbf{X}_t = (S_t, I_t, R_t)$.

2.4 Bayesian Inference

In a Bayesian framework, we are interested in the posterior distribution

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})}{\pi(\boldsymbol{y})},$$

where $\boldsymbol{\theta}$ are the unknown parameters, $\boldsymbol{y}$ is the vector of observed data, $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood of the observed data given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta})$ is the prior probability on $\boldsymbol{\theta}$. For complicated or high-dimensional models, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ may not have a closed form. However, using a Markov chain Monte Carlo (MCMC) approach, a Markov chain with stationary distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can be constructed.

2.4.1 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] samples unknown variables from a proposal distribution $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$, and accepts the proposed parameters $\boldsymbol{\theta}^*$ with probability equal to $\min(1, A)$, where

$$A = \frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta})} \frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})} \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}^*)}{\pi(\boldsymbol{y}|\boldsymbol{\theta})}$$

is the acceptance ratio. If the proposed values are accepted, $\boldsymbol{\theta}^*$ becomes the next value of the Markov chain. If they are rejected, the next step of the chain is equal to the previous step, $\boldsymbol{\theta}$. The stationary distribution of this Markov chain is the target distribution, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Thus, the Metropolis-Hastings algorithm allows us to sample from the target distribution using a simple, arbitrary proposal distribution. This can be extended to hidden Markov models or any other model with latent variables or missing data, where we are interested in the posterior

$$\pi(\boldsymbol{\theta}, \boldsymbol{X}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{y})}{\pi(\boldsymbol{y})} = \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})}{\pi(\boldsymbol{y})},$$

where $\boldsymbol{X}$ represents the latent variables, $\pi(\boldsymbol{X}|\boldsymbol{\theta})$ is the probability of the latent variables given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$ is the likelihood of the data given parameters $\boldsymbol{\theta}$ and latent states $\boldsymbol{X}$.

2.4.2 Particle marginal Metropolis-Hastings algorithm

A limitation of the Metropolis-Hastings algorithm is that the acceptance ratio requires the likelihood of the model using the proposed parameter values, and sometimes the likelihood for a hidden Markov model is computationally intractable. There exist likelihood free

methods that allow use of an approximation of the likelihood in the Metropolis-Hastings algorithm. One such method, the particle marginal Metropolis-Hastings (PMMH) algorithm, introduced by Beaumont [2003] and studied by Andrieu and Roberts [2009] and Andrieu et al. [2010], constructs a Markov chain that targets the joint posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{X}|\mathbf{y})$, where $\boldsymbol{X}$ is a set of auxiliary or hidden variables, and requires only an unbiased estimate of the likelihood. For an accessible exposition of the algorithm see [Wilkinson, 2011, Chapter 10].

The PMMH algorithm has two parts: an SMC algorithm, which is used to estimate the marginal likelihood of the data given a particular set of parameters, $\boldsymbol{\theta}$, and a Metropolis-Hastings step [Metropolis et al., 1953, Hastings, 1970], which uses the estimated likelihood in the acceptance ratio. The PMMH algorithm targets the joint posterior distribution $\pi(\boldsymbol{\theta}, \mathbf{X}|\mathbf{y})$ by jointly updating $\boldsymbol{\theta}$ and $\boldsymbol{X}_{1:T}$. At each step, a new $\boldsymbol{\theta}^*$ is proposed from the proposal distribution $q(\cdot|\boldsymbol{\theta})$. An SMC algorithm is used to generate and weight K particle trajectories of $\boldsymbol{X}_{1:T}$ corresponding to the hidden state processes using the proposed parameter set $\boldsymbol{\theta}^*$. A proposed $\boldsymbol{X}^*_{1:T}$ trajectory is sampled from the K trajectories ($\boldsymbol{X}^k_{1:T}$, for $k = 1, \ldots, K$) based on the final set of particle weights. The marginal likelihood is estimated by summing the weights of the SMC algorithm, and the proposed $\boldsymbol{\theta}^*$ and $\boldsymbol{X}^*_{1:T}$ are accepted with probability equal to the familiar Metropolis-Hastings acceptance ratio. The algorithm, which follows closely Andrieu et al. [2010], is detailed below.

Step 1: for iteration $i = 0$.

(a) Set $\boldsymbol{\theta}(0)$ arbitrarily

(b) Run an SMC algorithm to get $\hat{p}_{\boldsymbol{\theta}(0)}(\boldsymbol{y})$, an estimate of the marginal likelihood, and to produce a sample $\boldsymbol{X}_{1:T}(0) \sim \hat{p}_{\boldsymbol{\theta}(0)}(\cdot|\boldsymbol{y})$.

Step 2: for iteration $i \geq 1$,

(a) Sample $\boldsymbol{\theta}^* \sim q\{\cdot|\boldsymbol{\theta}(i-1)\}$

(b) Run an SMC algorithm to get $\hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})$ and $\boldsymbol{X}^*_{1:T} \sim \hat{p}_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{y})$

(c) with probability

$$\min\left\{1, \frac{\hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})}{\hat{p}_{\boldsymbol{\theta}(i-1)}(\boldsymbol{y})} \frac{\Pr(\boldsymbol{\theta}^*)}{\Pr\{\boldsymbol{\theta}(i-1)\}} \frac{q\{\boldsymbol{\theta}(i-1)|\boldsymbol{\theta}^*\}}{q\{\boldsymbol{\theta}^*|\boldsymbol{\theta}(i-1)\}}\right\}$$

set $\boldsymbol{\theta}(i) = \boldsymbol{\theta}^*$, $\boldsymbol{X}_{1:T}(i) = \boldsymbol{X}^*_{1:T}$, and $\hat{p}_{\boldsymbol{\theta}(i)}(\boldsymbol{y}) = \hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})$, otherwise set $\boldsymbol{\theta}(i) = \boldsymbol{\theta}(i-1)$, $\boldsymbol{X}_{1:T}(i) = \boldsymbol{X}_{1:T}(i-1)$, and $\hat{p}_{\boldsymbol{\theta}(i)}(\boldsymbol{y}) = \hat{p}_{\boldsymbol{\theta}(i-1)}(\boldsymbol{y})$

To construct an unbiased estimate of the likelihood, we use a sequential Monte Carlo (SMC) algorithm, also known as a bootstrap particle filter [Doucet et al., 2001]. The SMC algorithm sequentially estimates the likelihood using weighted particles; it requires the ability to propagate the unobserved data, $\boldsymbol{X}_t$, forward in time and the calculation of the probability of the observed data given the simulated unobserved data. Using a generic hidden Markov model, this probability can be expressed as

$$y_n|\mathbf{X}_n = \mathbf{x} \sim g_\theta(\cdot|\mathbf{x})$$

for time $n \in \{1, \ldots, T\}$. To propagate the latent variables forward in time, we sample from the initial density $\mathbf{X}_1 \sim \mu_\theta(\cdot)$ and sample from the transition probability density,

$$\mathbf{X}_{n+1}|\mathbf{X}_n = \mathbf{x} \sim f_\theta(\cdot|\mathbf{x}).$$

The following exposition of the SMC algorithm follows closely the pseudocode of Andrieu et al. [2010]. The superscript $k \in \{1, \ldots, K\}$ denotes the particle index and the subscript $n \in \{1, \ldots, T\}$ denotes the time; thus, $\boldsymbol{X}^k_{1:n}$ denotes the kth particle at time n.
Step 1: for $n = 1$,

(a) For $k = 1, \ldots, K$, sample

$$\begin{aligned} \boldsymbol{X}^k_1 &\sim q_\theta(\cdot|y_1) && \text{[Importance density]} \\ &= \mu_\theta(\boldsymbol{x}_1) && \text{[Initial density of hidden Markov state process in our case]} \end{aligned}$$

(b) For $k = 1, \ldots, K$,

$$\begin{aligned} w_1(\boldsymbol{X}_1^k) &:= \frac{p_\theta(\boldsymbol{X}_1^k, y_1)}{q_\theta(\boldsymbol{X}_1^k | y_1)} \qquad \text{[Importance weight]} \\ &= \frac{\mu_\theta(\boldsymbol{X}_1^k) g_\theta(y_1 | \boldsymbol{X}_1^k)}{q_\theta(\boldsymbol{X}_1^k | y_1)} \\ &= g_\theta(y_1 | \boldsymbol{X}_1^k), \\ W_1^k &:= \frac{w_1(\boldsymbol{X}_1^k)}{\sum_{m=1}^K w_1(\boldsymbol{X}_1^m)}. \qquad \text{[Normalize weights]} \end{aligned}$$

Step 2: for $n = 2, ..., T$,

(a) Sample $A_{n-1}^k \sim \mathcal{F}(\cdot | \mathbf{W}_{n-1})$ for $k = 1, \ldots, K$, where $\mathbf{W}_n := (W_n^1, ..., W_n^K)$ denotes the normalized importance weights at time n, and $\mathcal{F}(\cdot | \mathbf{p})$ denotes a discrete distribution with probability mass function $\mathbf{p}$. Here A_{n-1}^k denotes the index s of the 'parent' particle $X_{1:n-1}^s$ which is propagated forward at time n.

(b) For $k = 1, \ldots, K$, sample

$$\begin{aligned} \boldsymbol{X}_n^k &\sim q_\theta(\cdot | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k}) \qquad \text{[Importance density]} \\ &= f_\theta(\cdot | \boldsymbol{X}_{n-1}^{A_{n-1}^k}) \qquad \text{[Transition probability density in our case]} \end{aligned}$$

and set $\boldsymbol{X}_{1:n}^k := (\boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{X}_n^k)$.

(c) For $k = 1, \ldots, K$,

$$\begin{aligned}
w_n(\boldsymbol{X}_{1:n}^k) &:= \frac{p_\theta(\boldsymbol{X}_{1:n}^k, \boldsymbol{y}_{1:n})}{p_\theta(\boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{y}_{1:n-1}) q_\theta(\boldsymbol{X}_n^k | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k})} \\
&= \frac{p_\theta(\boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{X}_n^k, \boldsymbol{y}_{1:n-1}, y_n)}{p_\theta(\boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{y}_{1:n-1}) q_\theta(\boldsymbol{X}_n^k | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k})} \\
&= \frac{p_\theta(y_n, \boldsymbol{X}_n^k | \boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{y}_{1:n-1})}{q_\theta(\boldsymbol{X}_n^k | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k})} \\
&= \frac{p_\theta(y_n | \boldsymbol{X}_{1:n}^k, \boldsymbol{y}_{1:n-1}) p_\theta(\boldsymbol{X}_n^k | \boldsymbol{X}_{1:n-1}^{A_{n-1}^k}, \boldsymbol{y}_{1:n-1})}{q_\theta(\boldsymbol{X}_n^k | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k})} \\
&= \frac{f_\theta(\boldsymbol{X}_n^k | \boldsymbol{X}_{1:n-1}^{A_{n-1}^k}) g_\theta(y_n | \boldsymbol{X}_n^k)}{q_\theta(\boldsymbol{X}_n^k | y_n, \boldsymbol{X}_{n-1}^{A_{n-1}^k})} \qquad \text{[From the properties of HMMs]} \\
&= g_\theta(y_n | \boldsymbol{X}_n^k), \qquad \text{[Since we chose } q_\theta(\cdot | y_n, \boldsymbol{X}_n) = f_\theta(\cdot | \boldsymbol{X}_n)\text{]} \\
W_n^k &= \frac{w_n(\boldsymbol{X}_{1:n}^k)}{\sum_{m=1}^K w_n(\boldsymbol{X}_{1:n}^m)}. \qquad \text{[Normalize weights]}
\end{aligned}$$

Note that:

$$\begin{aligned}
p_\theta(y_n | \boldsymbol{y}_{1:n-1}) &= \int p_\theta(y_n, \boldsymbol{X}_{1:n} | \boldsymbol{y}_{1:n-1}) \mathrm{d}\boldsymbol{X}_{1:n} \\
&= \int \frac{p_\theta(y_n, \boldsymbol{X}_{1:n} | \boldsymbol{y}_{1:n-1})}{p_\theta(\boldsymbol{X}_{1:n-1} | \boldsymbol{y}_{1:n-1})} p_\theta(\boldsymbol{X}_{1:n-1} | \boldsymbol{y}_{1:n-1}) \mathrm{d}\boldsymbol{X}_{1:n} \\
&= \int \frac{p_\theta(\boldsymbol{y}_{1:n}, \boldsymbol{X}_{1:n})}{p_\theta(\boldsymbol{X}_{1:n-1}, \boldsymbol{y}_{1:n-1})} p_\theta(\boldsymbol{X}_{1:n-1} | \boldsymbol{y}_{1:n-1}) \mathrm{d}\boldsymbol{X}_{1:n} \\
&= \int p_\theta(y_n | \boldsymbol{X}_{1:n}, \boldsymbol{y}_{1:n-1}) p_\theta(\boldsymbol{X}_n | \boldsymbol{X}_{1:n-1}, \boldsymbol{y}_{1:n-1}) p_\theta(\boldsymbol{X}_{1:n-1} | \boldsymbol{y}_{1:n-1}) \mathrm{d}\boldsymbol{X}_{1:n} \\
&= \int w_n(\boldsymbol{X}_{1:n}) q_\theta(\boldsymbol{X}_n | y_n, \boldsymbol{X}_{n-1}) p_\theta(\boldsymbol{X}_{1:n-1} | \boldsymbol{y}_{1:n-1}) \mathrm{d}\boldsymbol{X}_{1:n}.
\end{aligned}$$

From above, it follows that

$$\hat{p}_\theta(y_n | \boldsymbol{y}_{1:n-1}) = \frac{1}{K} \sum_{k=1}^K w_n(\boldsymbol{X}_{1:n}^k) \quad \underset{\text{a.s.}}{\to} \quad p_\theta(y_n | \boldsymbol{y}_{1:n-1})$$

as $K \to \infty$, and

$$\hat{p}_\theta(\boldsymbol{y}) = \hat{p}_\theta(y_1) \prod_{n=2}^{T} \hat{p}_\theta(y_n | \boldsymbol{y}_{1:n-1}).$$

Thus we have a sequential, likelihood-free algorithm which generates an unbiased estimate of the marginal likelihood, $p_{\boldsymbol{\theta}}(\boldsymbol{y})$. The PMMH algorithm uses this estimated marginal likelihood in the acceptance ratio of the Metropolis Hastings algorithm.

Chapter 3

SIRS ANALYSIS OF CHOLERA OUTBREAKS IN MATHBARIA, BANGLADESH

3.1 Introduction

In Bangladesh, cholera is an endemic disease that demonstrates seasonal outbreaks [Huq et al., 2005, Koelle and Pascual, 2004, Koelle et al., 2005, Longini et al., 2002]. The burden of cholera is high in that country, with an estimated 352,000 cases and 3,500 to 7,000 deaths annually [International Vaccine Institute, 2012]. We seek to understand the dynamics of cholera and to develop a model that will be able to predict outbreaks several weeks in advance. If the timing and size of a seasonal epidemic could be predicted reliably, vaccines and other resources could be allocated effectively to curb the impact of the disease.

Specifically, we want to understand how the disease dynamics are related to environmental covariates. It is currently not known what triggers the seasonal cholera outbreaks in Bangladesh, but it has been shown that *Vibrio cholerae*, the causative bacterial agent of cholera, can be detected in the environment year round [Huq et al., 1990, Colwell and Huq, 1994]. Environmental forces are thought to contribute to the spread of cholera, evident from the many cholera disease dynamics models that incorporate the role of the aquatic environment on cholera transmission through an environmental reservoir effect [Codeço, 2001, Tien and Earn, 2010]. One hypothesis is that proliferation of *V. cholerae* in the environment triggers the seasonal epidemic, feedback from infected individuals drives the epidemic, and then cholera outbreaks wane, either due to an exhaustion of the susceptibles or due to the deteriorating ecological conditions for propagation of *V. cholerae* in the environment. We probe this hypothesis using cholera incidence data and ecological data collected from multiple thanas in rural Bangladesh over sixteen years. There have been three phases of data collection so far, each lasting approximately three years and being separated by gaps of a few years; the current collection phase is ongoing. For a subset of these data, Huq et al.

[2005] used Poisson regression to study the association between lagged predictors from a particular water body to cholera cases in that thana. This resulted in different lags and different significant covariates across multiple water bodies and thanas. Thus, it was hard to derive a cohesive model for predicting cholera outbreaks from the environmental covariates. Also, there is no easy way to account for disease dynamics in this Poisson regression framework. We want to measure the effect of the environmental covariates while accounting for disease dynamics via mechanistic models of disease transmission. Moreover, we want to see if we can make reliable short-term predictions with our model — a task that was not attempted by Huq et al. [2005].

Mechanistic infectious disease models use scientific understanding of the transmission process to develop dynamical systems that describe the evolution of the process [Bretó et al., 2009]. Realistic models of disease transmission incorporate non-linear dynamics [He et al., 2010], which leads to difficulty in the statistical inference under these models, specifically in the tractability of the likelihood. Keeling and Ross [2008] demonstrate some of these difficulties; they use an exact stochastic continuous-time, discrete-state model which evolves Markov processes using the deterministic Kolmogorov forward equations to express the probabilities for being in all possible states. However, that method only works for small populations due to computational limitations. To overcome this intractability, Finkenstädt and Grenfell [2000] develop a time-series Susceptible-Infected-Recovered (SIR) model which extends mechanistic models of disease dynamics to larger populations. A similar development in analysis of infectious disease time-series is the auto-Poisson model of Held et al. [2005]. To facilitate tractability of the likelihood, both of the above approaches make simplifying assumptions that are difficult to test. Moreover, these discrete-time approaches work only for evenly spaced data or require aggregating the data into evenly spaced intervals. Cauchemez and Ferguson [2008] develop a different, continuous-time, approach to analyze epidemiological time-series data, but assume the transmission parameter and number of susceptibles remain relatively constant within an observation period. Our current understanding of cholera disease dynamics makes us think that this assumption is not appropriate for modeling endemic cholera with seasonal outbreaks.

To implement a mechanistic approach without these approximations, both maximum

likelihood and Bayesian methods could be used. Maximum likelihood based statistical inference techniques use Monte Carlo to allow maximization of the likelihood without explicitly evaluating it [He et al., 2010, Bretó et al., 2009, Ionides et al., 2006, Bhadra et al., 2011]. Ionides et al. [2006] use this methodology to study how large scale climate fluctuations influence cholera transmission in Bangladesh. Bhadra et al. [2011] use this framework to study malaria transmission in India. They are able to incorporate a rainfall covariate into their model and study how climate fluctuations influenced disease incidence when one controlled for disease dynamics, such as waning immunity. Using a Bayesian approach, particle filter Markov chain Monte Carlo (MCMC) methods have been developed which require only an unbiased estimate of the likelihood [Andrieu et al., 2010]. Rasmussen et al. [2011] use this particle MCMC methodology to simultaneously estimate the epidemiological parameters of a SIR model and past disease dynamics from time series data and gene genealogies. Using Google flu trends data [Ginsberg et al., 2008], Dukic et al. [2012] implement a particle filtering algorithm which sequentially estimates the odds of a pandemic. Notably, Dukic et al. [2012] concentrate on predicting influenza activity. Similarly, here we develop a model-based predictive framework for seasonal cholera epidemics in Bangladesh.

In this paper, we use sequential Monte Carlo methods in a Bayesian framework. Specifically, we develop a hidden Susceptible-Infected-Recovered-Susceptible (SIRS) model for cholera transmission in Bangladesh, incorporating environmental covariates. We use a particle MCMC method to sample from the posterior distribution of the environmental and transmission parameters given the observed data, as described by Andrieu et al. [2010]. Further, we predict future behavior of the epidemic within our Bayesian framework. Cholera transmission dynamics in our model are described by a continuous-time, rather than a discrete-time, Markov process to easily incorporate data with irregular observation times. Also, the continuous-time framework allows for greater parameter interpretability and comparability to models based on deterministic differential equations. We test our Bayesian inference procedure using simulated cholera data, generated from a model with a time-varying environmental covariate. We then analyze cholera data from Mathbaria, Bangladesh, similar to the data studied by Huq et al. [2005]. Parameter estimates indicate that most of the transmission is coming from environmental sources. We test the ability of our model to

make short-term predictions during different time points in the data observation period and find that the pattern of predictive distribution dynamics matches the pattern of changes in the reported number of cases. Moreover, we find that the predictive distribution of the hidden states, specifically the unobserved number of infected individuals, clearly pinpoints the beginning of an epidemic approximately two to three weeks in advance, making our methodology potentially useful during cholera surveillance in Bangladesh.

3.2 SIRS model with environmental predictors

We consider a compartmental model of disease transmission [May and Anderson, 1991, Keeling and Rohani, 2008], where the population is divided into three disease states, or compartments: susceptible, infected, and recovered. We model a continuous process observed at discrete time points. The vector $\boldsymbol{X}_t = (S_t, I_t, R_t)$ contains the numbers of susceptible, infected, and recovered individuals at time t, and we consider a closed population of size N such that $N = S_t + I_t + R_t$ for all t. Individuals move between the compartments with different rates; for cholera transmission we consider the transition rates shown in Figure 3.1. In this framework, a susceptible individual's rate of infection is proportional to the number of infected people and the covariates that serve as proxy for the amount of *V. cholerae* in the environment. Thus, the hazard rate of infection, also called the force of infection, is $\beta I_t + \alpha(t)$ for each time t, where β represents the infectious contact rate between infected individuals and susceptible individuals and $\alpha(t)$ represents the time-varying environmental force of infection. Possible mechanisms for infectious contact include direct person-to-person transmission of cholera and consumption of water that has been contaminated by infected individuals. Infected individuals recover from infection at a rate γ, where $1/\gamma$ is the average length of the infectious period. Once the infected individual has recovered from infection, they move to the recovered compartment. Recovered individuals develop a temporary immunity to the disease after infection. They move from the recovered compartment to the susceptible compartment with rate μ, where $1/\mu$ is the average length of immunity. Birth and death are incorporated into the system indirectly through the waning of immunity; thus, instead of just representing natural loss of immunity, μ also represents the loss of immunity through the death of recovered individuals and birth of new susceptible individuals.

The models of Codeço [2001] and Koelle and Pascual [2004] similarly allow susceptibles to be renewed at a rate that combines net birth and loss of immunity.

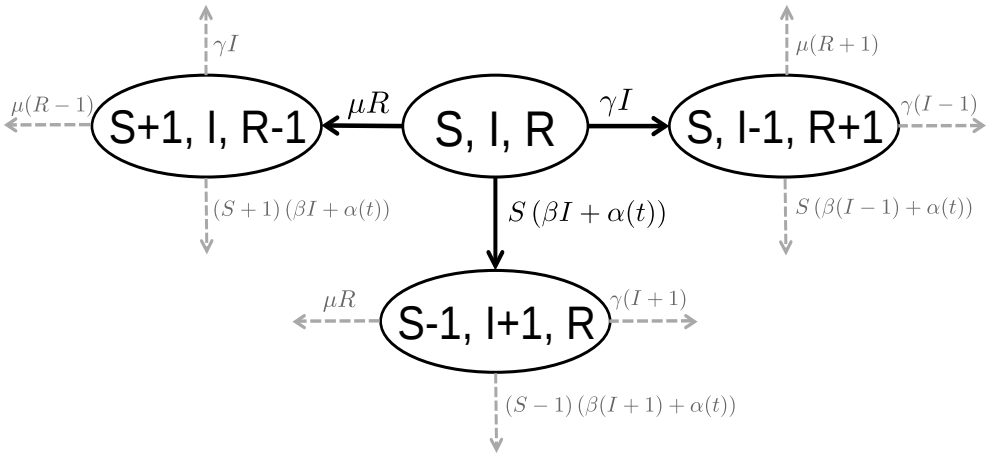

Figure 3.1: State transitions for Susceptible-Infected-Recovered-Susceptible (SIRS) model for cholera. S, I, and R denote the numbers of susceptible, infected, and recovered individuals. The state can transition to one of three new states. These new states correspond to a susceptible becoming infected, an infected recovering from infection, or a recovered individual losing immunity to infection and becoming susceptible. The parameter β is the infectious contact rate, $\alpha(t)$ is the time-varying environmental force of infection, γ is the recovery rate, and μ is the rate at which immunity is lost.

We model $\boldsymbol{X}_t$ as an inhomogeneous Markov process [Taylor and Karlin, 1998] with infinitesimal rates

$$\lambda_{(S,I,R),(S',I',R')}(t) = \begin{cases} (\beta I + \alpha(t))\, S & \text{if } S' = S-1,\ I' = I+1,\ R' = R, \\ \gamma I & \text{if } S' = S,\ I' = I-1,\ R' = R+1, \\ \mu R & \text{if } S' = S+1,\ I' = I,\ R' = R-1, \\ 0 & \text{otherwise,} \end{cases} \tag{3.1}$$

where $\boldsymbol{X} = (S, I, R)$ is the current state and $\boldsymbol{X}' = (S', I', R')$ is a new state. Because $R_t = N - S_t - I_t$, we keep track of only susceptible and infected individuals, S_t and I_t.

This type of compartmental model is similar to other cholera models in the literature. The time-series SIRS model of Koelle and Pascual [2004] also includes the effects of both intrinsic factors (disease dynamics) and extrinsic factors (environment) on transmission. King et al. [2008] examines both a regular SIRS model and a two-path model to include asymptomatic infections, and uses a time-varying transmission term that incorporates transmission via the environmental reservoir and direct person-to-person transmission, but does not allow for feedback from infected individuals into the environmental reservoir. The SIWR model of Tien and Earn [2010] and Eisenberg et al. [2013b] allows for infections from both a water compartment (W) and direct transmission and considers the feedback created by infected individuals contaminating the water. To allow for the possibility of asymptomatic individuals, Longini et al. [2007] use a model with a compartment for asymptomatic infections; that model only considers direct transmission. Codeço [2001] uses an SIR model with no direct person-to-person transmission; infected individuals excrete directly into the environment and susceptible individuals are infected from exposure to contaminated water. Our SIRS model is not identical to any of the above models, but it borrows from them two important features: explicit modeling of disease transmission from either direct person-to-person transmission of cholera or consumption of water that has been contaminated by infected individuals and a time-varying environmental force of infection.

3.3 Hidden SIRS model

While the underlying dynamics of the disease are described by $\boldsymbol{X}_t$, these states are not directly observed. The number y_t of infected individuals observed at each time point t is only a random fraction of the number of infected individuals. This fraction depends on both the number of infected individuals that are symptomatic and the fraction of symptomatic infected individuals that seek treatment and get reported (the reporting rate). Thus, y_{t_i}, the number of observed infections at time t_i for observation $i \in \{0, 1, ..., n\}$, has a binomial distribution with size I_{t_i}, the number of infected individuals at time t_i, and success probability ρ, the probability of infected individuals seeking treatment, so

$$\Pr(y_{t_i}|I_{t_i}, \rho) = \binom{I_{t_i}}{y_{t_i}} \rho^{y_{t_i}} (1-\rho)^{I_{t_i} - y_{t_i}}. \tag{3.2}$$

Given $\boldsymbol{X}_{t_i}$, y_{t_i} is independent of the other observations and other hidden states.

We use a Bayesian framework to estimate the parameters of the hidden SIRS model, where the unobserved states $\boldsymbol{X}_t$ are governed by the infinitesimal rates in Equation (3.1). The parameters that we want to estimate are β, γ, μ, ρ, and the $k+1$ parameters that will be incorporated into $\alpha(t)$, the time-varying environmental force of infection. We assume $\alpha(t) = \exp\left(\alpha_0 + \alpha_1 C_1(t) + \cdots + \alpha_k C_k(t)\right)$, where $C_1(t), \ldots, C_k(t)$ denote the k time-varying environmental covariates.

We assume independent Poisson initial distributions for S_{t_0} and I_{t_0}, with means ϕ_S and ϕ_I respectively. Thus $\Pr(\boldsymbol{X}_{t_0}|\phi_S, \phi_I) = \Pr(S_{t_0}|\phi_S) \times \Pr(I_{t_0}|\phi_I) = \frac{\phi_S^{S_{t_0}} \exp(-\phi_S)}{S_{t_0}!} \times \frac{\phi_I^{I_{t_0}} \exp(-\phi_I)}{I_{t_0}!}$. Parameters that are constrained to be greater than zero, such as β, γ, μ, ϕ_S, and ϕ_I, are transformed to the log scale. A logit transformation is used for the probability ρ. We assume independent normal prior distributions on all of the transformed parameters, incorporating biological information into the priors where possible.

We are interested in the posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{y}) \propto \Pr(\boldsymbol{y}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta})$, where $\boldsymbol{y} = (y_{t_0}, \ldots, y_{t_n})$, $\boldsymbol{\theta} = (\log(\beta), \log(\gamma), \log(\mu), \text{logit}(\rho), \alpha_0, \ldots, \alpha_k, \log(\phi_S), \log(\phi_I))$, and

$$\Pr(\mathbf{y}|\boldsymbol{\theta}) \;=\; \sum_{\boldsymbol{x}} \left(\prod_{i=0}^{n} \Pr(y_{t_i}|I_{t_i}, \rho) \left[\Pr(\boldsymbol{X}_{t_0}|\phi_S, \phi_I) \prod_{i=1}^{n} p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_{i-1}}, \boldsymbol{\theta}) \right] \right).$$

Here $p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_{i-1}}, \boldsymbol{\theta})$ for $i = 1, \ldots, n$ are the transition probabilities of the continuous-time Markov chain (CTMC). However, this likelihood is intractable; there is no practical method to compute the finite time transition probabilities of the SIRS CTMC because the size of the state space of $\boldsymbol{X}_t$ grows on the order of N^2. For the same reason, summing over $\boldsymbol{X}$ with the forward-backward algorithm [Baum et al., 1970] is not feasible. To use Bayesian inference despite this likelihood intractability, we turn to a particle marginal Metropolis-Hastings (PMMH) algorithm [Andrieu et al., 2010], which uses a sequential Monte Carlo technique to generate an estimate of the likelihood using simulations of the unobserved states [Doucet et al., 2001].

3.4 Particle filter MCMC

3.4.1 Overview

The particle marginal Metropolis-Hastings algorithm, introduced by Beaumont [2003] and studied in [Andrieu and Roberts, 2009, Andrieu et al., 2010], constructs a Markov chain that targets the joint posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{X}|\mathbf{y})$, where $\boldsymbol{X}$ is a set of auxiliary or hidden variables, and requires only an unbiased estimate of the likelihood. To construct this likelihood estimate, we use a sequential Monte Carlo (SMC) algorithm, also known as a bootstrap particle filter [Doucet et al., 2001]. The SMC algorithm sequentially estimates the likelihood using weighted particles; it requires the ability to propagate the unobserved data, $\boldsymbol{X}_t$, forward in time and the calculation of the probability of the observed data given the simulated unobserved data. For the hidden SIRS model, $y_{t_i}|\boldsymbol{X}_{t_i} = (S_{t_i}, I_{t_i}, R_{t_i}) \sim \text{Binomial}(I_{t_i}, \rho)$, where ρ depends on the number of symptomatic infected individuals that seek treatment, as described in Section 3.3. Thus the probability of the observed data given the simulated unobserved data is given by Equation (3.2). To propagate the hidden variables forward in time, we first simulate initial states $\boldsymbol{X}_{t_0} = (S_{t_0}, I_{t_0})$ from Poisson distributions with means ϕ_S and ϕ_I respectively. We then use properties of CTMCs to simulate the trajectories of the unobserved states.

Thus, the PMMH algorithm has two parts: an SMC algorithm, which is used to estimate the marginal likelihood of the data given a particular set of parameters, $\boldsymbol{\theta}$, and a Metropolis-

Hastings step [Metropolis et al., 1953, Hastings, 1970], which uses the estimated likelihood in the acceptance ratio. At each step, a new $\boldsymbol{\theta}^*$ is proposed from the proposal distribution $q(\cdot|\boldsymbol{\theta})$. An SMC algorithm is used to generate and weight K particle trajectories corresponding to the hidden state processes using the proposed parameter set $\boldsymbol{\theta}^*$. A proposed $\boldsymbol{X}^*_{\boldsymbol{t}_{0:n}}$ trajectory is sampled from the K particle trajectories based on the final particle weights of the SMC algorithm. The marginal likelihood is estimated by summing the weights of the SMC algorithm, and the proposed $\boldsymbol{\theta}^*$ and $\boldsymbol{X}^*_{\boldsymbol{t}_{0:n}}$ are accepted with probability equal to the familiar Metropolis-Hastings acceptance ratio.

3.4.2 PMMH pseudocode

The following exposition of the algorithm follows closely the pseudocode of Andrieu et al. [2010].

Step 1: initialization, for iteration $j = 0$,

(a) Set $\boldsymbol{\theta}(0)$ arbitrarily

(b) Run the following SMC algorithm to get $\hat{p}_{\boldsymbol{\theta}(0)}(\boldsymbol{y})$, an estimate of the marginal likelihood, and to produce a sample $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(0) \sim \hat{p}_{\boldsymbol{\theta}(0)}(\cdot|\boldsymbol{y})$.

Let the superscript $k \in \{1, \ldots, K\}$ denote the particle index, where K is the total number of particles, and the subscript $t_i \in \{t_0, \ldots, t_n\}$ denote the time; thus, $\boldsymbol{X}^k_{t_i}$ denotes the kth particle at time t_i, and $\boldsymbol{X}^k_{\boldsymbol{t}_{0:i}} = \left(\boldsymbol{X}^k_{t_0}, \ldots, \boldsymbol{X}^k_{t_i}\right)$. At time $t_i = t_0$, sample $\boldsymbol{X}^k_{t_0}$ for $k = 1, \ldots, K$ from the initial density of the hidden Markov state process. Specifically, sample $S_{t_0} \sim \text{Poisson}(\phi_S)$ and $I_{t_0} \sim \text{Poisson}(\phi_I)$. Compute the k weights $w(\boldsymbol{X}^k_{t_0}) := \Pr(y_{t_0}|\boldsymbol{X}^k_{t_0})$.

For $i = 1, \ldots, n$, sample particles with probabilities proportional to their weights and propagate resampled particles forward. Combine the trajectory associated with the resampled particle and simulated next particle to define the $\boldsymbol{X}^k_{\boldsymbol{t}_{0:i}}$ trajectory. Continue computing weights $w(\boldsymbol{X}^k_{t_i}) := \Pr(y_{t_i}|\boldsymbol{X}^k_{t_i})$ and propagating resampled particles forward until $i = n$.

It follows that

$$\hat{p}_{\boldsymbol{\theta}}(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}}) = \frac{1}{K}\sum_{k=1}^{K} w(\boldsymbol{X}_{t_i}^k)$$

is an approximation to the likelihood $p_{\boldsymbol{\theta}}(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}})$, and therefore an approximation to the total likelihood is

$$\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \hat{p}_{\boldsymbol{\theta}}(y_{t_0})\prod_{i=1}^{n}\hat{p}_{\boldsymbol{\theta}}(y_{t_i}|\boldsymbol{y}_{\boldsymbol{t}_{0:i-1}}).$$

Thus we have a simple, sequential, likelihood-free algorithm which generates an unbiased estimate of the marginal likelihood, $p_{\boldsymbol{\theta}}(\boldsymbol{y})$. A $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(0)$ trajectory is sampled from the K trajectories ($\boldsymbol{X}^k_{\boldsymbol{t}_{0:n}}$, for $k = 1, \ldots, K$) based on the final set of particle weights.

Step 2: for iteration $j \geq 1$,

(a) Sample $\boldsymbol{\theta}^* \sim q\{\cdot|\boldsymbol{\theta}(j-1)\}$

(b) Run an SMC algorithm, as in step 1(b) with $\boldsymbol{\theta}^*$ instead of $\boldsymbol{\theta}(0)$, to get $\hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})$ and $\boldsymbol{X}^*_{\boldsymbol{t}_{0:n}} \sim \hat{p}_{\boldsymbol{\theta}^*}(\cdot|\boldsymbol{y})$

(c) With probability

$$\min\left\{1, \frac{\hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})}{\hat{p}_{\boldsymbol{\theta}(j-1)}(\boldsymbol{y})}\frac{\Pr(\boldsymbol{\theta}^*)}{\Pr\{\boldsymbol{\theta}(j-1)\}}\frac{q\{\boldsymbol{\theta}(j-1)|\boldsymbol{\theta}^*\}}{q\{\boldsymbol{\theta}^*|\boldsymbol{\theta}(j-1)\}}\right\}$$

set $\boldsymbol{\theta}(j) = \boldsymbol{\theta}^*$, $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j) = \boldsymbol{X}^*_{\boldsymbol{t}_{0:n}}$, and $\hat{p}_{\boldsymbol{\theta}(j)}(\boldsymbol{y}) = \hat{p}_{\boldsymbol{\theta}^*}(\boldsymbol{y})$, otherwise set $\boldsymbol{\theta}(j) = \boldsymbol{\theta}(j-1)$, $\boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j) = \boldsymbol{X}_{\boldsymbol{t}_{0:n}}(j-1)$, and $\hat{p}_{\boldsymbol{\theta}(j)}(\boldsymbol{y}) = \hat{p}_{\boldsymbol{\theta}(j-1)}(\boldsymbol{y})$.

To propagate the unobserved $\boldsymbol{X}_t = (S_t, I_t, R_t)$ forward in time, we simulate from a cholera transmission model with a time-varying environmental force of infection. CTMCs which incorporate time-varying transition rates are inhomogeneous. The details of the discretely-observed inhomogeneous CTMC simulations are in Section 3.4.3.

3.4.3 Simulating inhomogeneous SIRS using tau-leaping

Much work has been done on simulating continuous-time, discrete space Markov processes, especially in the area of chemical kinetics. Gillespie developed two methods for exact

stochastic simulation of chemical process trajectories with constant rates: the direct method [Gillespie, 1977] and the first reaction method [Gillespie, 1976]. Details of these methods are given in Chapter 2.

The exact algorithms work for small populations, but for large state spaces these methods require a prohibitively long computing time. This is a common problem in chemical kinetics literature, where an approximate method called the tau-leaping algorithm originated [Gillespie, 2001, Cao et al., 2005]. This method simulates CTMCs by jumping over a small amount of time τ and approximating the number of events that happen in this time using a series of Poisson distributions. As τ approaches zero, this approximation theoretically approaches the exact algorithm. The value of τ must be chosen such that the rates remain roughly constant over the period of time; this is referred to as the "leap condition".

Specifically, for our simulation, using the methods outlined in Cao et al. [2005], we define the rate functions $h_1(\boldsymbol{X}_t) = (\beta I_t + \alpha(t))\, S_t$, $h_2(\boldsymbol{X}_t) = \gamma I_t$, and $h_3(\boldsymbol{X}_t) = \mu R_t$, corresponding to the infinitesimal rates of the CTMC. Then $k_1 \sim \text{Poisson}(h_1(\boldsymbol{X}_t)\tau)$ represents the number of infections in time $[t, t+\tau)$, $k_2 \sim \text{Poisson}(h_2(\boldsymbol{X}_t)\tau)$ represents the number of recoveries in time $[t, t+\tau)$, and $k_3 \sim \text{Poisson}(h_3(\boldsymbol{X}_t)\tau)$ represents the number of people that become susceptible to infection in time $[t, t+\tau)$. We make the assumption that the time-varying force of infection, $\alpha(t)$, remains constant each day. We define daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, and $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$. Using $\tau = 1$ day, our rates now remain constant within each tau jump. To see if this value for τ is reasonable, we perform a simulation study. Using the posterior estimates of the parameters, we simulate data forward in time 5000 times using both the modified Gillespie algorithm and the modified tau-leaping algorithm. We simulate data over the entire epidemic curve to see how the comparison changes for varying values of $\alpha(t)$. Figure 3.2 shows estimates of the median and 95% intervals for the simulated values. The Monte Carlo standard error is very small for all estimates. For the numbers of susceptible individuals, the estimates under Gillespie and tau-leaping are almost identical over the entire epidemic. For the numbers of infected, the values are very close except at the epidemic peaks. However, the differences are very small. We conclude that for our application $\tau = 1$ day is a good compromise between computational efficiency and accuracy.

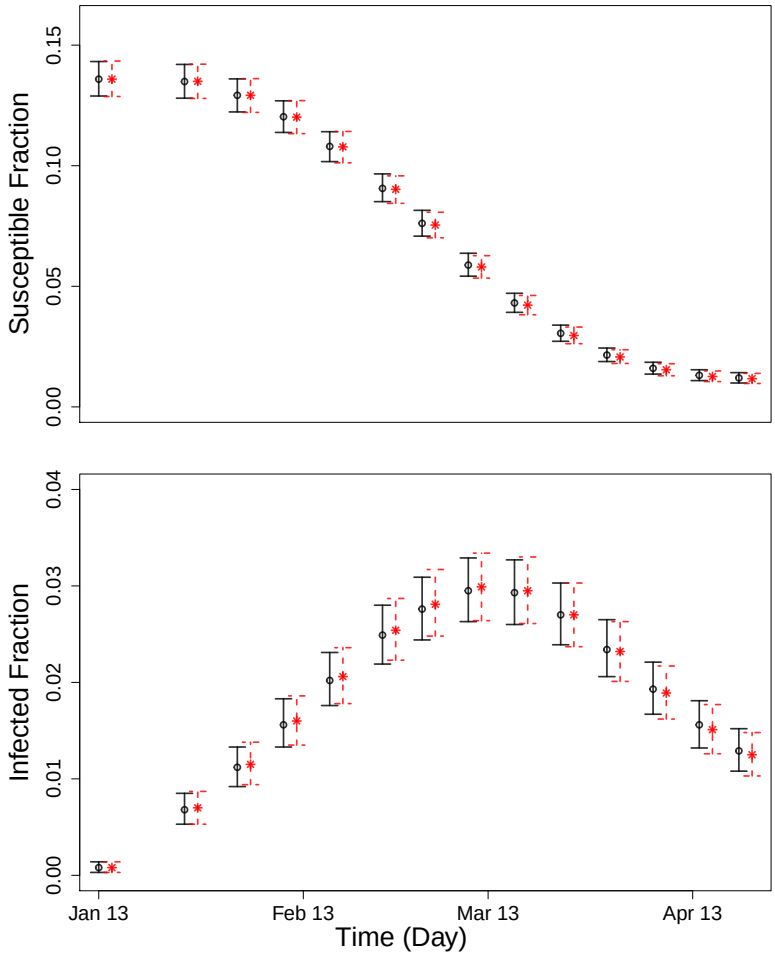

Figure 3.2: Plots comparing the median and 95% intervals at different points during an epidemic, simulated using both the modified Gillespie algorithm and the modified tau-leaping algorithm with $\tau = 1$ day. The medians and 95% intervals for 5000 simulations using the Gillespie algorithm are given by the open circle and solid error bars, respectively. The medians and 95% intervals for 5000 simulations using the modified tau-leaping algorithm are given by the asterisk and dashed error bars, respectively.

3.4.4 Metropolis-Hastings proposal for model parameters

Our implementation of the PMMH algorithm starts with a preliminary run, which consists of a burn-in run plus a secondary run, both using independent normal proposal distributions for the parameters. From the secondary run, we calculate the approximate posterior covariance of the parameters and use it to construct the covariance of the multivariate normal proposal

distribution in the final run of the PMMH algorithm. In all runs, parameters are proposed and updated jointly.

3.4.5 Prediction

One of the main goals of this analysis is to be able to predict cholera outbreaks in advance using environmental predictors. To assess the predictive ability of our model, we estimate the parameters of the model using a training set of data and then predict future behavior of the epidemic process. We examine the posterior predictive distributions of cholera counts by simulating data forward in time under the time-varying SIRS model using the accepted parameter values explored by the particle MCMC algorithm and the accepted values of the hidden states S_T and I_T at the final observation time, $t = T$, of the training data. These hidden states are sampled in the PMMH algorithm by sampling the last set of particles using the last set of weights [Andrieu et al., 2010]. Under each set of parameters, we generate possible future hidden states and observed data, and we compare the posterior predictive distribution of observed cholera cases to the test data. In the analyses below, the PMMH output is always thinned to 500 iterations for prediction purposes by saving only every kth iteration, where k depends on the total number of iterations.

3.5 Simulation results

To test the PMMH algorithm on simulated infectious disease data, we generate data from a hidden SIRS model with a time-varying environmental force of infection. We then use our Bayesian framework to estimate the parameters of the simulated model and compare the posterior distributions of the parameters with the true values. To simulate endemic cholera where many people have been previously infected, we start with a population size $N = 10000$ and assume independent Poisson initial distributions for S_{t_0} and I_{t_0}, with means $\phi_S = 2100$ and $\phi_I = 15$. The other parameters are set at $\beta = 1.25 \times 10^{-5}$, $\gamma = 0.1$, and $\mu = 0.0009$. All the rates are measured in the number of events per day. The average length of the infectious period, $1/\gamma$, is set to be 10 days, and the average length of immunity, $1/\mu$, is set to be about 3 years. Parameter values are chosen such that the simulated data is similar to the data collected from Mathbaria, Bangladesh. We use the daily time intervals

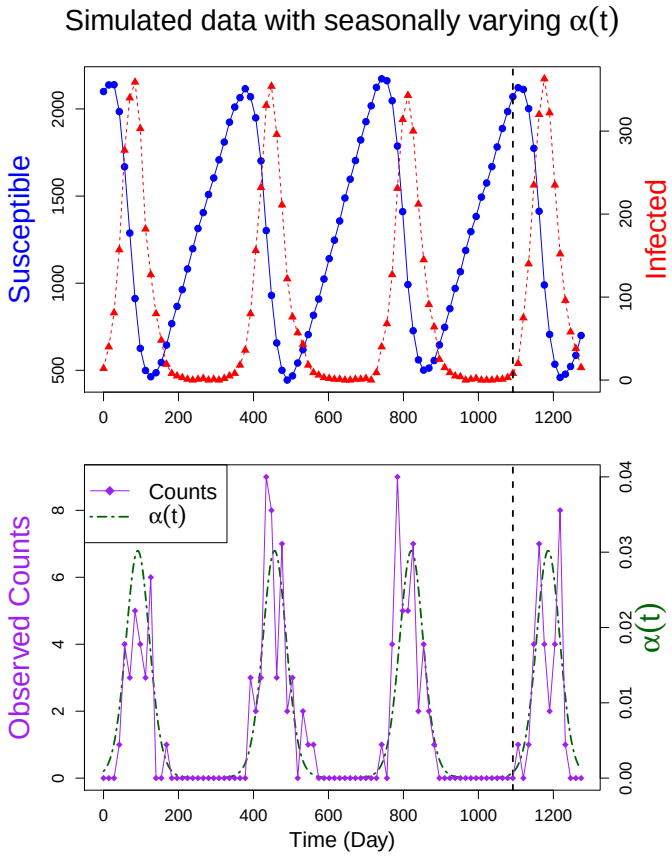

Figure 3.3: Plots of simulated hidden states (counts of susceptible, S_t, and infected, I_t, individuals) and the observed data ($y_t \sim \text{Binomial}(I_t, \rho)$ = number of observed infections) plotted over time (t). Simulation with seasonally varying $\alpha(t)$ generates data with seasonal epidemic peaks. The dashed vertical black line represents the first cut off between the training sets and the test data. Data before the line was used to estimate parameters, and we use those estimates to predict the data after the line. Other data cut offs are shown in Figure 3.7

$A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, defined in Section 3.4.3, and define $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$ where $\alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 \sin\left(\frac{2\pi}{365} i\right)\right]$. The intercept α_0 and the amplitude α_1 are parameters to be estimated. The frequency of the sine function is set to mimic the annual peak seen in the environmental data collected from Bangladesh. For the data simulation, we set $\alpha_0 = -7$ and $\alpha_1 = 3.5$. Using the modified Gillespie algorithm described in Section 2.2.2, we simulate the (S_t, I_t) chain given in the top plot of Figure 3.3. The observed number of infections $y_t \sim \text{Binomial}(I_t, \rho)$, where $\rho = 0.015$ and is treated as an unknown parameter.

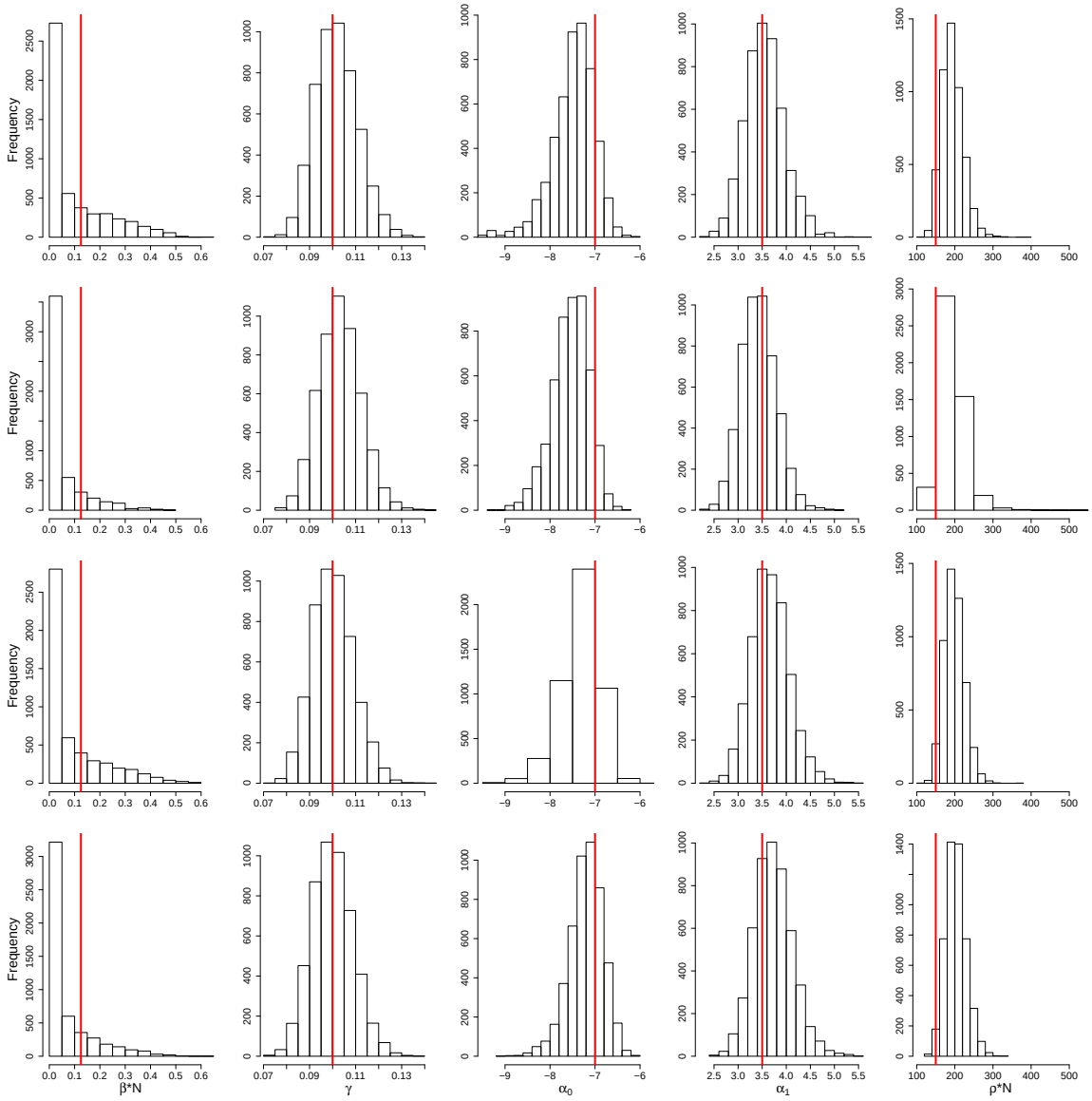

Figure 3.4: Posterior distributions for the parameters of the SIRS model with time-varying environmental force of infection. From top to bottom, assumed ϕ_S/N and ϕ_I/N are above the true values, (0.31 and 0.003), at the truth (0.21 and 0.0015), below the true values (0.11 and 0.00075), and further below (0.055 and 0.000375). The true values of the parameters are denoted by the red lines in the posterior histograms. Posterior distributions look similar regardless of assumptions about ϕ_S/N and ϕ_I/N. However, the range of the posterior for $\beta \times N$ is lowest and the range of the posterior for $\rho \times N$ is highest when ϕ_S/N and ϕ_I/N are assumed to be larger than the true values.

We simulate three years of training data because this is approximately how long the data collection phases last in our data from Bangladesh [Huq et al., 2005]. Therefore we do not attempt to estimate the loss of immunity rate μ since it is on the scale of three years. Also, there is not enough information in the data to estimate the means of the Poisson initial distributions, ϕ_S and ϕ_I, since estimation of these parameters is only informed by the very beginning of the observed data. We set these parameters to different values and compare parameter estimation and prediction between models with parameter assumptions which differ from the truth. We also assume $N = 10000$.

We assume normal prior distributions on all of the transformed parameters, with means and standard deviations chosen so that the true value of the parameter is in the tail of the prior distribution. We use relatively uninformative, diffuse priors for $\log(\beta)$, α_0, and α_1, centered at $\log(1.25 \times 10^{-4})$, -8, and 0, respectively, and with standard deviations of 5. The prior distribution for $\text{logit}(\rho)$ is centered at $\text{logit}(0.03)$ and has a standard deviation of 2. For $\log(\gamma)$, the prior is centered at $\log(0.1)$ with a standard deviation of 0.09, since this value is well studied for cholera. Thus, the 95% range for the prior on $1/\gamma$ is 8.38 to 11.93 days.

Using these data, the PMMH algorithm starts with a burn-in run of 10000 iterations, a secondary run of 10000 iterations, and a final run of 50000 iterations. To thin the chains, we save only every 10th iteration. We use $K = 100$ particles in the SMC algorithm. We compare results from models with different assumptions on the values of ϕ_S and ϕ_I: assumed ϕ_S and ϕ_I are above the true values (0.31 and 0.003), at the true values (0.21 and 0.0015), below the true values (0.11 and 0.00075), or further below the true values (0.055 and 0.000375). Marginal posterior distributions for the parameters of the SIRS model from the final runs of these PMMH algorithms are in Figure 3.4. The posterior distributions look similar regardless of the assumed values of ϕ_S and ϕ_I, though estimates for $\beta \times N$ and $\rho \times N$ deviate furthest from the truth when ϕ_S and ϕ_I are set higher than the true values. Trace plots and auto-correlation plots for the parameters of the SIRS model from the final run of the PMMH algorithm for simulated data are in Figure 3.5, and Figure 3.6 shows bivariate scatterplots of the parameters. We report $\beta \times N$ and $\rho \times N$, since we found these to be robust to assumptions about the total population size N in sensitivity analyses. Effective

sample sizes, a measure of the number of independent samples in a Markov chain, for the posterior samples range from 593 to 2038 for the parameters of the SIRS model assuming the true values of ϕ_S and ϕ_I.

From the posterior distributions, it is clear that the algorithm is providing good estimates of the true parameter values. Estimates of the parameter $\beta \times N$ are slightly different than the true value, causing the estimate of the reporting rate to be a little off as well. This parameter relationship makes sense, as a high ρ accounts for the low number of infected individuals resulting from a small infectious contact rate β.

3.5.1 Prediction results

To test the predictive ability of the model, we use multiple cut off times to separate our simulated data into staggered training sets and test sets. The complete data are shown in the bottom row of Figure 3.3. For each cut off time, parameters were drawn from the posterior distribution based on the training data. These parameter values were then used to simulate possible realizations of reported infections after the training data until the next cut off 28 days later. The distributions of these predicted reported cases are shown in the top plot of Figure 3.7. The test data are denoted by the purple diamonds, connected by straight lines to help visualize ups and downs in the case counts. Case counts are observed once every 14 days. On each observation day, the colored bar represents the distribution of predicted counts for that day. As desired, the posterior predictive distribution shifts its mass as time progresses to follow the case counts in the test data. The plot of the predicted hidden states in the bottom row of Figure 3.7 also shows that our model is capturing the formation and decline of the epidemic peak well, as seen in the trajectory of the predicted fraction of infected individuals. This plot illustrates the interplay of the hidden states of the underlying compartmental model. During an epidemic, the fraction of susceptibles decreases while the fraction of infected individuals quickly increases. Afterwards, the fraction of infected individuals drops and the pool of susceptibles slowly begins to increase as both immunity is lost and more susceptible individuals are born.

These predictions were made under the assumption that ϕ_S and ϕ_I are set to the true

values. To test sensitivity to these assumptions, we compare predictions made from models that assume other values; these are shown in Figure 3.8. Values for ϕ_S/N and ϕ_I/N are set over the true values, (0.31, 0.003), below the true values (0.11, 0.00075), and further below the truth (0.055, 0.000375). Predicted distributions look similar for all values of ϕ_S and ϕ_I. Uncertainty is greatest when ϕ_S and ϕ_I are set at higher values than the truth. For the lowest values of ϕ_S and ϕ_I, the fraction of susceptible individuals is lower and the fraction of infected is higher than those predicted fractions under other settings. However, important information, like the timing of the epidemic, remains intact.

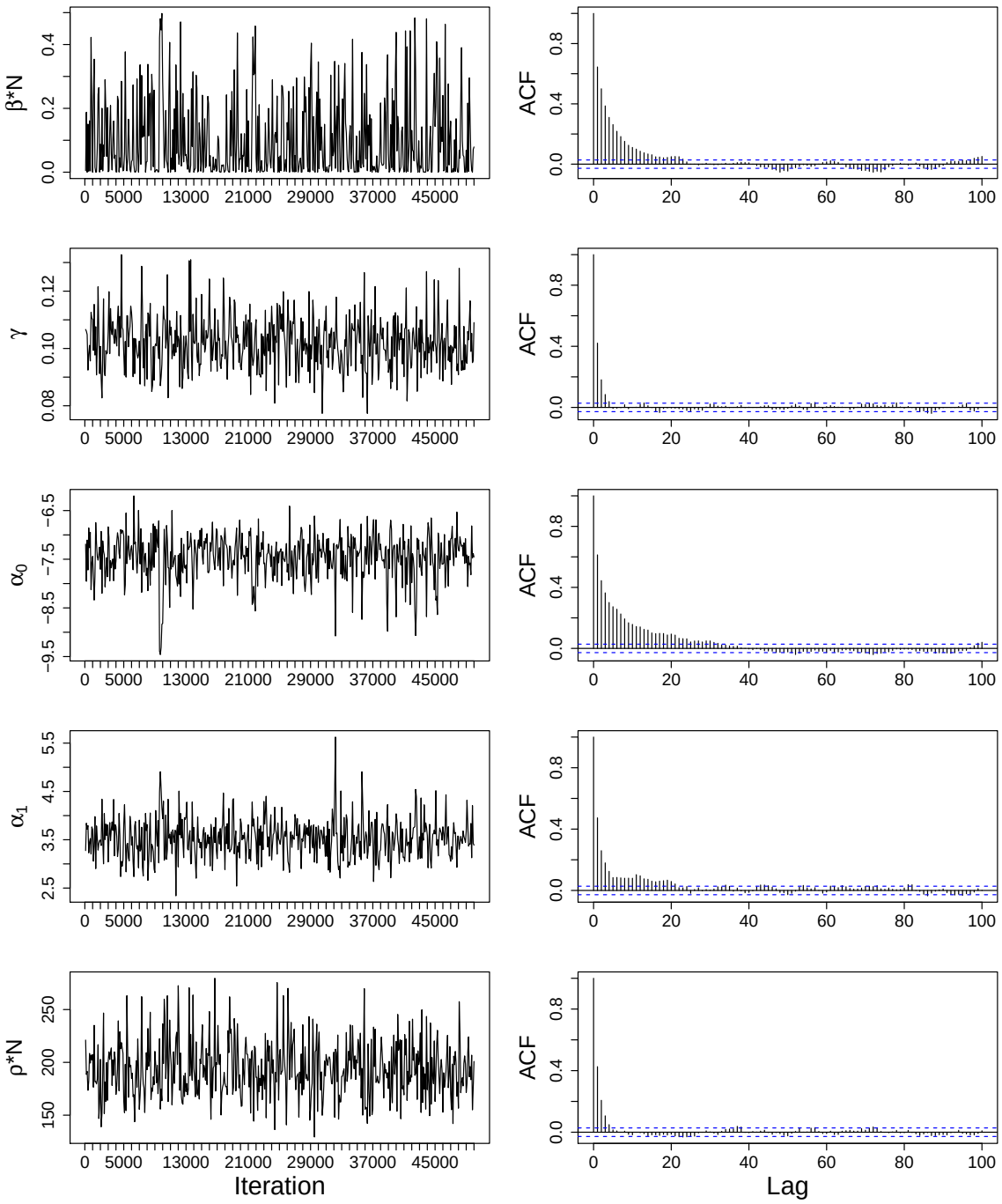

Figure 3.5: Summary plots of the PMMH algorithm output (final run of 50000 iterations) for the parameters of the SIRS model with a time-varying environmental force of infection. ACF plots are thinned to 5000 iterations and trace plots are thinned to display only 500 iterations.

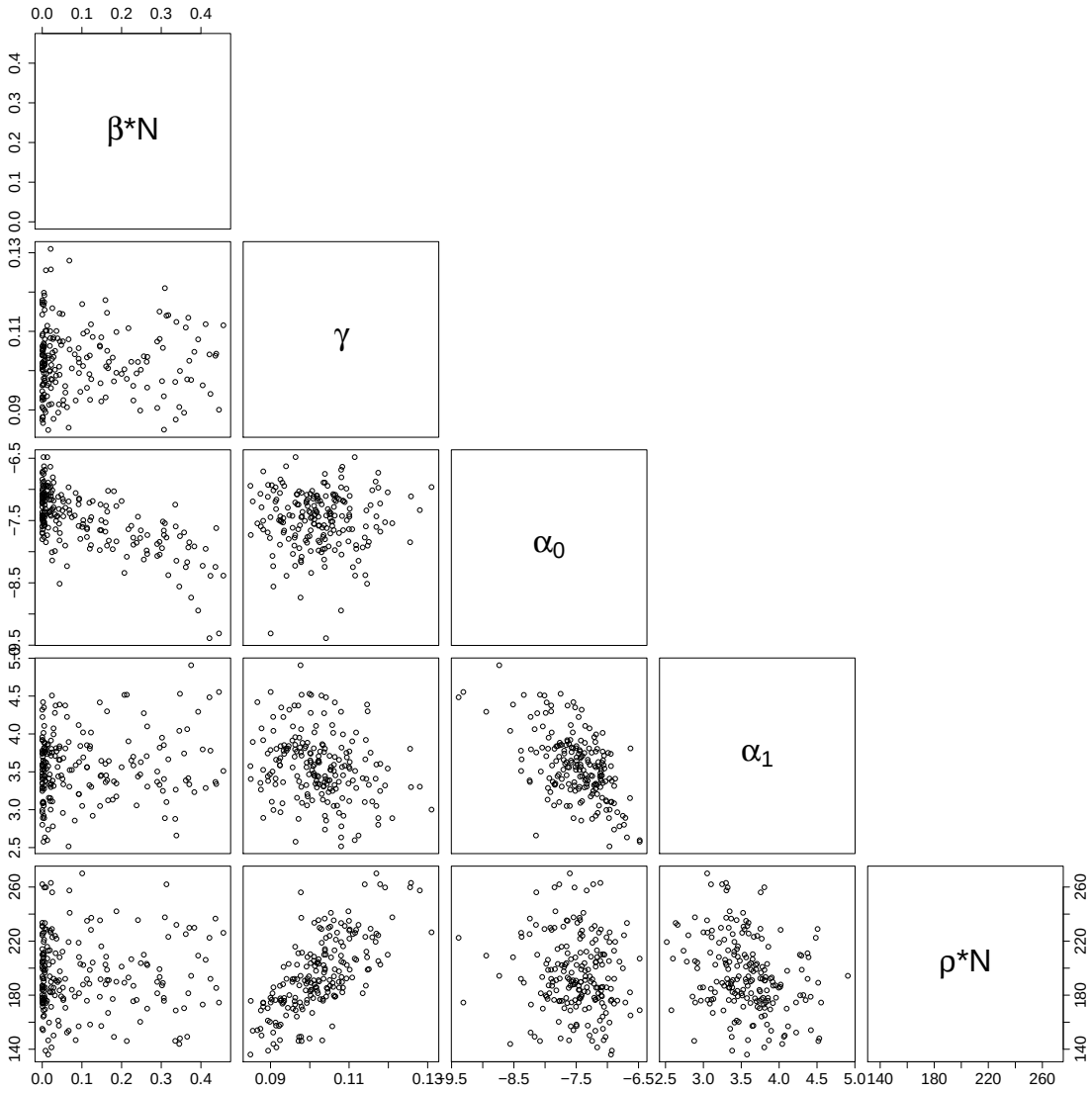

Figure 3.6: Bivariate scatterplots of parameters of the SIRS model with a time-varying environmental force of infection. Scatterplots are thinned to display only 200 samples, so only every 25th sample from the posterior distribution is plotted.

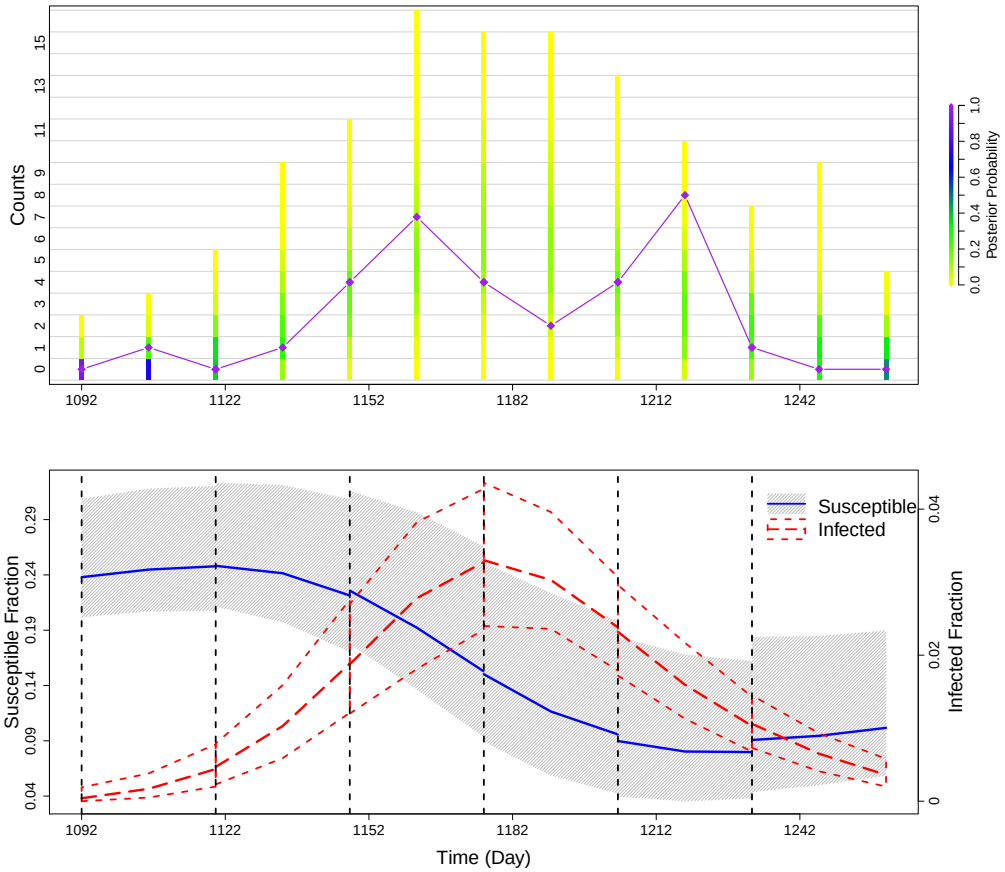

Figure 3.7: Summary of prediction results for simulated data. We run PMMH algorithms on training sets of the data, which are cut off at each of the dashed black lines in the bottom plot. Future cases are then predicted until the next cut off. The top plot compares the posterior probability of the predicted counts to the test data (diamonds connected by straight purple line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The bottom plot shows how the trajectory of the predicted hidden states changes over the course of the epidemic. The gray area and the solid line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of infected individuals.

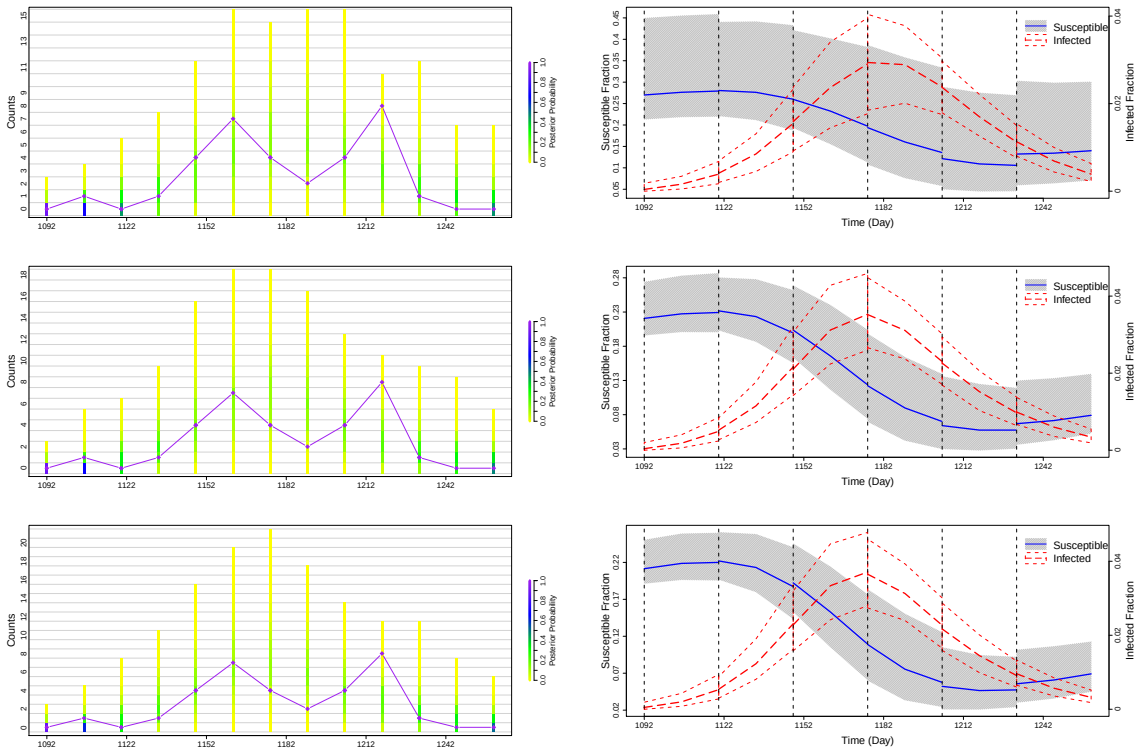

Figure 3.8: Summary of prediction results for simulated data; each row shows prediction results under different assumptions about the values of ϕ_s/N and ϕ_I/N. From top to bottom, values for ϕ_S/N and ϕ_I/N are set over the true values, (0.31, 0.003), below the true values (0.11, 0.00075), and further below (0.055, 0.000375). Plots on the left compare the posterior probability of the predicted counts to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The plots on the right show how the trajectory of the predicted hidden states change over the course of the epidemic. The gray area and the solid line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of infected individuals. Predictions look similar for all settings, though there is greater uncertainty in the predicted fraction of susceptibles when ϕ_S/N and ϕ_I/N are set higher than the true values; quantile ranges decrease as ϕ_S/N and ϕ_I/N decrease. The timing of the epidemic is the same for all settings.

3.6 Using cholera incidence data and covariates from Mathbaria, Bangladesh

Huq et al. [2005] found that water temperature (WT) and water depth (WD) in some water bodies had a significant lagged relationship with cholera incidence. Therefore, we use these covariates and cholera incidence data from Mathbaria, Bangladesh collected between April 2004 to September 2007 (Phase 2 of data collection) and again from October 2010 to July 2013 (Phase 3). During these time periods, cholera incidence data were collected over a period of three days approximately every two weeks. Environmental data were also collected approximately every two weeks from six water bodies. To get a smooth summary of the covariates using data from all water bodies, we fit a cubic spline to the covariate values. We then slightly modify our environmental force of infection to allow for a lagged covariate effect. Let κ denote the length of the lag. We consider the daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$ and define the environmental force of infection $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$ and $t \geq \kappa$ where $\alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 C_{WD}(i-\kappa) + \alpha_2 C_{WT}(i-\kappa)\right]$. Here the covariates are the smoothed standardized daily values $C_{WT}(i) = (WT(i) - \overline{WT}_P)/s_{WT,P}$ and $C_{WD}(i) = (WD(i) - \overline{WD}_P)/s_{WD,P}$, where $\overline{X}_P$ is the mean of the measurements for all i in Phase $P \in \{\text{Phase 2, Phase 3}\}$ and $s_{X,P}$ is the sample standard deviation. We consider and compare results from models assuming three different lags: $\kappa = 14$, $\kappa = 18$, and $\kappa = 21$. Predictions from all three models are similar, so we report only results from the model assuming $\kappa = 21$ in order to receive the earliest warning of upcoming epidemics; see Figures 3.13 and 3.14 for prediction comparisons. The smoothed, standardized, 21 day lagged covariates and cholera incidence data are shown in Figure 3.9.

Since there is only about six years of data, estimating the loss of immunity rate μ is infeasible. Thus, we set $\mu = 0.0009$ so that $1/\mu$ is 3 years [Sack et al., 2004]. Also, the population size N, which quantifies the size catchment area for the medical center, is assumed to be 10000 for computational convenience. We do not know the true value of N, but 10000 seems reasonable and is small enough that simulations run quickly. We studied sensitivity to these assumptions by setting both μ and N to different values, obtaining similar results. We also again set ϕ_S and ϕ_I to various values and obtained similar results. See Section 3.6.1 for details.

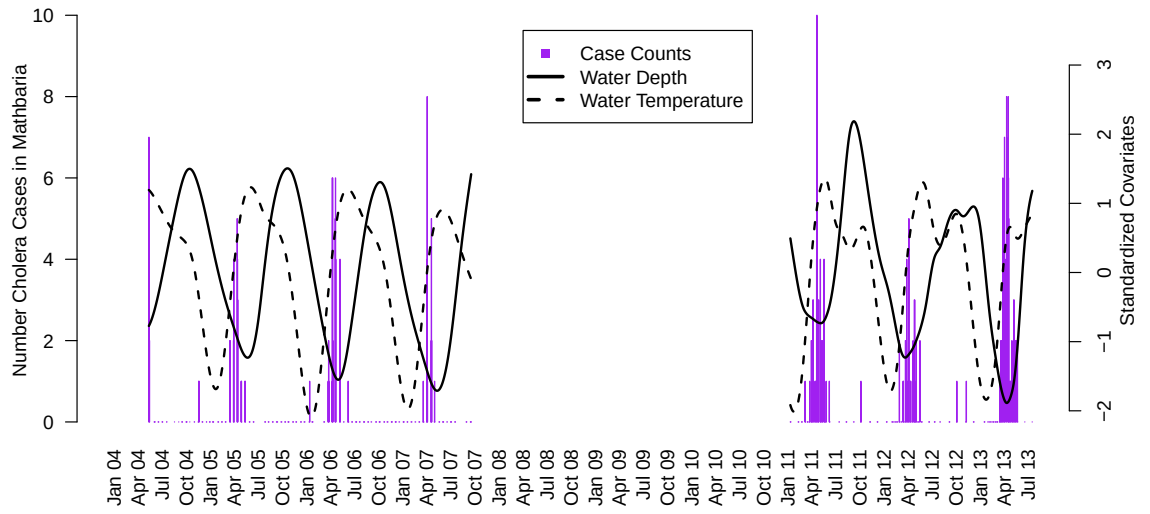

Figure 3.9: Barplot of cholera case counts in Mathbaria, Bangladesh and the standardized covariate measurements over time. The covariates are shown with a lag of three weeks. No data were collected from October 2007 through November 2010. The ranges of the unstandardized smoothed covariates are 1.42 to 2.76 meters for water depth and 21.59 to 33.12°C for water temperature.

In these runs, we use relatively uninformative, diffuse normal prior distributions on the time-varying environmental covariates α_1 and α_2, centered at 0 and with standard deviations of 5. The diffuse normal prior distributions on the transformed parameter values $\log(\beta)$ and α_0 are centered at $\log(1.25 \times 10^{-7})$ and -8, respectively, with standard deviations of 5. We know that the average infectious period for cholera, $1/\gamma$, should be between 8 and 12 days. Thus, the transformed parameter $\log(\gamma)$ is given a normal prior distribution with mean $\log(0.1)$ and standard deviation 0.09. We also know that ρ should be very close to zero, since only a small proportion of cholera infections are symptomatic and a smaller proportion will be treated at the health complex [Sack et al., 2003]. Thus, the transformed parameter $\text{logit}(\rho)$ is given a normal prior distribution with mean $\text{logit}(0.0008)$ and standard deviation equal to 2.

We implement the PMMH algorithm with a burn-in run of 30000 iterations, a secondary run of 20000 iterations, and a final run of 400000 iterations. We again save only every 10th

iteration and use $K = 100$ particles in the SMC algorithm. Posterior medians and 95% Bayesian credible intervals for the parameters $\beta \times N$, γ, α_0, α_1, α_2, and $\rho \times N$ generated by the final run of the PMMH algorithm are given in Table 3.1. We report $\beta \times N$ and $\rho \times N$ since we found these parameter estimates to be robust to changes in the population size N during sensitivity analyses. For more details, see Section 3.6.1. The credible intervals for α_1 and α_2 do not include zero, so both water depth and water temperature have a significant relationship with the force of infection. Decreasing water depth increases the force of infection, likely due to the higher concentration and resulting proliferation of *V. cholerae* in the environment; increasing water temperature increases the force of infection [Huq et al., 2005].

The reproductive number, R, is the average number of secondary cases caused by a typical infected individual in a population that consists of both susceptible and immune individuals [Diekmann et al., 1990]. We report $(\beta \times N)/\gamma$, the part of the reproductive number that is related to the number of infected individuals in the population under our model assumptions. The estimate is close to zero, suggesting that the epidemic peaks in our model are driven mostly by the environmental force of infection. However, the infectious contact rate is not zero and is not negligible compared to the environmental force of infection. In particular, posterior median values for $\alpha(t)$ range from 0.00005 to 0.4, while posterior median values for βI_t only range from 0 to 0.00035.

Table 3.1: Posterior medians and 95% equitailed credible intervals (CIs) for the parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh.

Coefficient	Estimate	95% CIs
$\beta \times N$	0.0046	(0 , 0.5178)
γ	0.10	(0.09 , 0.12)
$(\beta \times N)/\gamma$	0.04	(0 , 4.65)
α_0	-5.35	(-6.12, -4.75)
α_1	-1.56	(-1.98, -1.12)
α_2	1.98	(1.58 , 2.44)
$\rho \times N$	48.8	(39.6 , 60.31)

3.6.1 MCMC diagnostics and sensitivity analysis

Summary plots of the PMMH algorithm output for the parameters of the SIRS model with data from Mathbaria, Bangladesh are given in Figure 3.10, and Figure 3.11 shows bivariate scatterplots of the parameters. Effective sample sizes range from 141 to 5265 for the analysis of the data from Mathbaria. To test convergence, we varied the initial values for the parameters of the PMMH algorithm. Some of the initial values are shown in Table 3.2 and the parameter estimates from the chains that started at these initial values are given in the top third of Table 3.3.

In our analysis of the data from Mathbaria, we assumed the size of the population, $N = 10000$, the loss of immunity rate, $\mu = 0.0009$, and the means of the Poisson initial distributions $\phi_S = .2 \times N$ and $\phi_I = 0.02 \times N$, are known. We studied sensitivity to these assumptions by setting all of these parameters to different values, and the results are shown in the bottom two-thirds of Table 3.3. We report $\beta \times N$ and $\rho \times N$ since we found these parameter estimates to be robust to changes in the population size N. As seen in Table 3.3, estimates are similar over different values of N, μ, ϕ_S and ϕ_I.

We also tested the effect of incorrect values for ϕ_S and ϕ_I on prediction using simulated data, as seen in Figure 3.8.

3.6.2 Prediction Results

For the data collected from Mathbaria, we begin prediction at multiple points around the time of the two epidemic peaks that occur in 2012 and 2013. Figure 3.9 shows the full cholera data with smoothed and standardized covariates. Figure 3.12 shows the posterior predictive distribution of observed cholera cases (top row) and hidden states from the time-varying SIRS model (bottom row). Parameters used to simulate the SIRS forward in time have been sampled using the PMMH algorithm applied to the training data, with data being cut off at different points during the 2012 and 2013 epidemic peaks. From each of these cut offs, parameter values are then used to simulate possible realizations of the test data. Predictions are run until the next cut off point, with cut off points chosen based on the length of the lag κ. Realistically, at time t we have covariate information to use for prediction only until time

$t + \kappa$, where κ is the covariate lag. Since the smallest lag considered is 14 days, we make only 14 day ahead predictions where possible to mimic a realistic prediction set up. Due to the sparse sampling between epidemic peaks (June 2012 to February 2013), we use longer prediction intervals for these cut-offs than would be possible in real time data analysis in order to evaluate our model predictions.

In the top row of Figure 3.12, the coloring of the bars again represents the distribution of predicted cases. Between the two peaks of case counts (June 2012 to February 2013), the frequency of predicted zero counts is very high, so we conclude that the model is doing well with respect to predicting the lack of an epidemic. During the epidemics, the distribution of the counts shifts its mass away from zero. The plot in the bottom row of Figure 3.12 again illustrates the periodic nature and interplay of the hidden states of the underlying compartmental model. When the fraction of infected individuals quickly increases during an epidemic, the fraction of susceptibles decreases. Afterwards, the fraction of infected individuals drops to almost zero and the pool of susceptibles is slowly replenished. When the fraction of infected individuals is low, there is more uncertainty in the prediction for the fraction of susceptibles (September 2012 to March 2013). The fraction of infected individuals increases to a slightly higher epidemic peak 2013 (March 2013 to May 2013) than in 2012 (March 2012 to May 2012), as observed in the test data for those years. The predicted fraction of infected people in the population increases before an increase can be seen in the case counts, which could allow for early warning of an epidemic.

To select a lag for the environmental covariates in the Mathbaria analysis, we compare prediction results from models assuming three different lags: $\kappa = 14$, $\kappa = 18$, and $\kappa = 21$. These are shown in Figures 3.13 and 3.14. The predictive distributions of the hidden states look similar across lags, so we use the 21 day lag model in order to predict an upcoming epidemic furthest in advance. With a three week lag, we would be able to make predictions three weeks in advance.

We also use a quasi-Poisson regression model similar to the one used by Huq et al. [2005] for prediction of cholera cases. For the two predictors, water temperature (WT) and water

depth (WD), we have

$$\ln \mathrm{E}(Y_t|C_{WD}(t-\kappa), C_{WT}(t-\kappa)) = \beta_0 + \beta_1 C_{WD}(t-\kappa) + \beta_2 C_{WT}(t-\kappa),$$

where $\kappa = 21$ days. The quasi-Poisson model accounts for overdispersion in the data [McCullagh and Nelder, 1989]. Figure 3.15 shows the predicted means and 95% intervals under the quasi-Poisson model. Test data are again cut off at different points during the 2012 and 2013 epidemic peaks and predictions are run until the next cut off point, with cut off points chosen approximately every two weeks. Predicted mean number of reported cases from the hidden SIRS model are also shown for comparison. Both models predict well the timing of epidemic peaks. However, the SIRS predicted fraction of infected individuals — a hidden variable in the SIRS model — provides a more detailed picture of how cholera affects a population. By providing not only accurate prediction of the time of epidemic peaks, but also the predicted fraction of the population that is infected, the SIRS model predictions could be used for efficient resource allocation to treat infected individuals.

3.7 Discussion

We use a Bayesian framework to fit a nonlinear dynamical model for cholera transmission in Bangladesh which incorporates environmental covariate effects. We demonstrate these techniques on simulated data from a hidden SIRS model with a time-varying environmental force of infection, and the results show that we are recovering well the true parameter values. We also estimate the effect of two environmental covariates on cholera case counts in Mathbaria, Bangladesh while accounting for infectious disease dynamics, and we test the predictive ability of our model. Overall, the prediction results look promising. The predicted hidden states show a noticeable increase in the fraction of infected individuals weeks before the observed number of cholera cases increases, which could allow for early notification of an epidemic and timely allocation of resources. The predicted hidden states show that the fraction of infected individuals in the population decreases greatly between epidemics, supporting the hypothesis that the environmental force of infection triggers outbreaks. Estimates of βI_t are low, but not negligible, compared to estimates of $\alpha(t)$, suggesting that

most of the transmission is coming from environmental sources.

Computational efficiency is an important factor in determining the usefulness of this approach in the field. We have written an R package which implements the PMMH algorithm for our hidden SIRS model, available at `https://github.com/vnminin/bayessir`. The computationally expensive portions of the PMMH code are primarily written in C++ to optimize performance, using Rcpp to integrate C++ and R [Eddelbuettel and François, 2011, Eddelbuettel, 2013]; however there is still room for improvement. Running 400000 iterations of the PMMH algorithm on the six years of data from Mathbaria takes two days on a 4.3 GHz i7 processor. Since we can predict three weeks into the future using a 21 day covariate lag, we do not think timing is a big limitation for using our model predictions in real life.

Figure 3.16 shows plots of standardized residuals versus time for each of the two phases of data collection in Mathbaria, Bangladesh. Standardized residuals are calculated as $\epsilon_{t_i} = \left(y_{t_i} - \mathrm{E}(y_{t_i})\right) / \mathrm{sd}(y_{t_i})$, where y_{t_i} is the number of observed infections at time t_i for observation $i \in \{0, 1, ..., n\}$. $\mathrm{E}(y_{t_i})$ and $\mathrm{sd}(y_{t_i})$ are approximated via simulation by fixing the model parameters to the posterior medians, running the SIRS model forward 5000 times, and computing the average and sample standard deviation of the 5000 realizations of the case counts at each time point. The residuals in Figure 3.16 show that we are modeling well case counts between the epidemic peaks but not the epidemic peaks themselves, either due to missing the timing of the epidemic peak or the latent states not being modeled accurately. This possible model misspecification might be fixed by including more covariates, using different lags, or modifying the SIRS model. Also, we assume a constant reporting rate, ρ, rather than using a time-varying ρ_t [Finkenstädt and Grenfell, 2000]. With better quality data we might be able to allow for a reporting rate that varies over time; we will try to address these model refinements in future analyses.

In the future, we will extend this analysis to allow for variable selection over a large number of covariates. This will allow us to include many covariates at many different lags and incorporate information from all of the water bodies in a way that does not involve averaging. In the current PMMH framework, choosing an optimal proposal distribution to explore a much larger parameter space would be difficult. We want to include a way

of automatically selecting covariates or shrinking irrelevant covariate effects to zero with sparsity inducing priors. The particle Gibbs sampler, introduced by Andrieu et al. [2010], would allow for such extensions. Approximate Bayesian computation is also an option for further model development [McKinley et al., 2009]. In addition, the available data consist of observations from multiple thanas during the same time period. Future analyses will look into sharing information across space and time and accounting for correlations between thanas. Another challenging future direction involves exploring models which incorporate a feedback loop from infected individuals back into the environment to capture the effect of infected individuals excreting *V. cholerae* into the environment. To accomplish this, we could add a water compartment to our SIRS model that quantifies the concentration of *V. cholerae* in the environment, similar to the SIWR model of Tien and Earn [2010]. However, adding an additional latent state leads to identifiability problems, even with fully observed data [Eisenberg et al., 2013b], so such an extension will require rigorous testing and fine tuning.

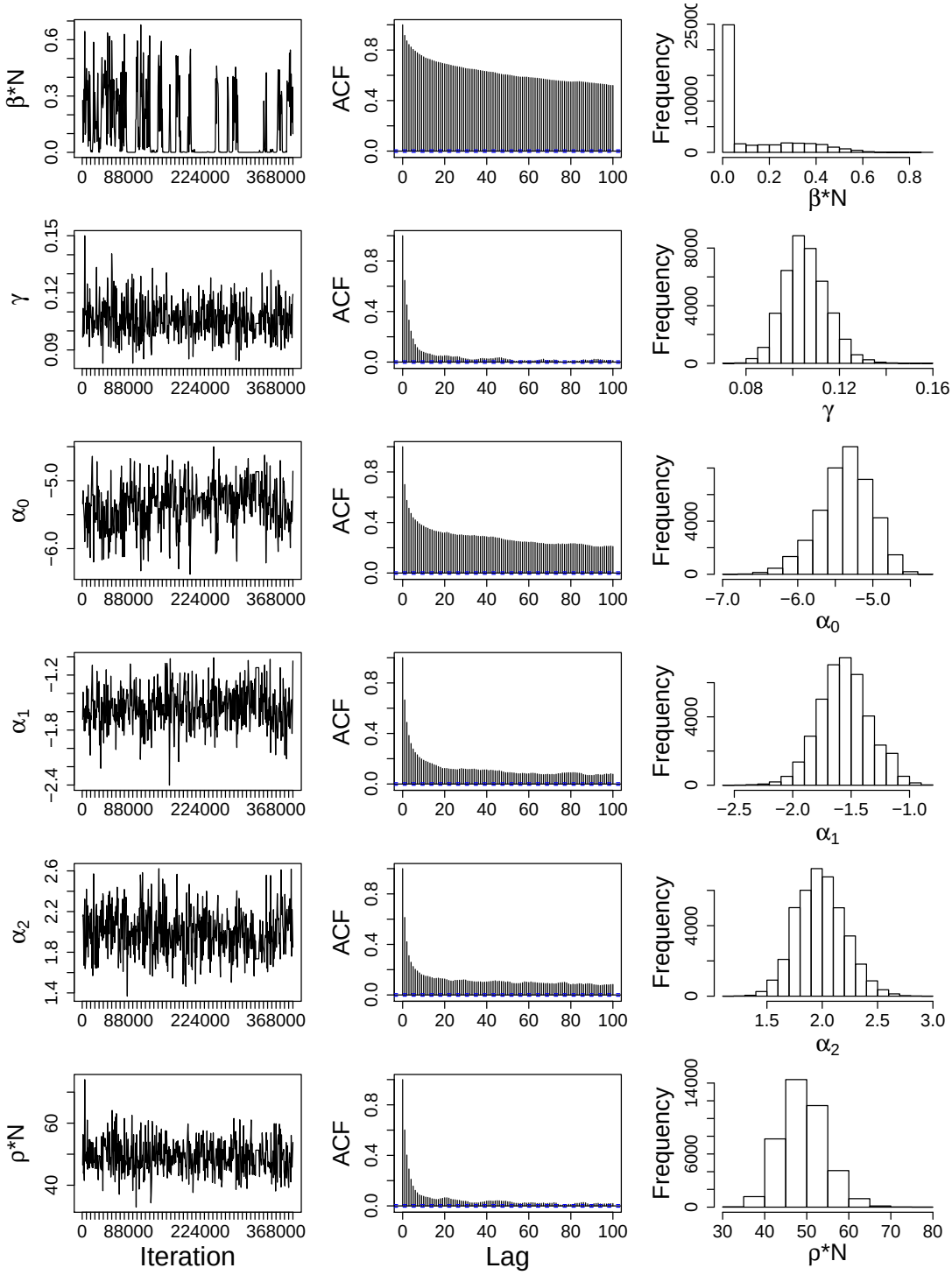

Figure 3.10: Summary plots of the PMMH algorithm output (final run of 400000 iterations) for the parameters of the SIRS model with data from Mathbaria, Bangladesh. ACF plots and histograms are thinned to 40000 iterations and trace plots are thinned to display only 500 iterations.

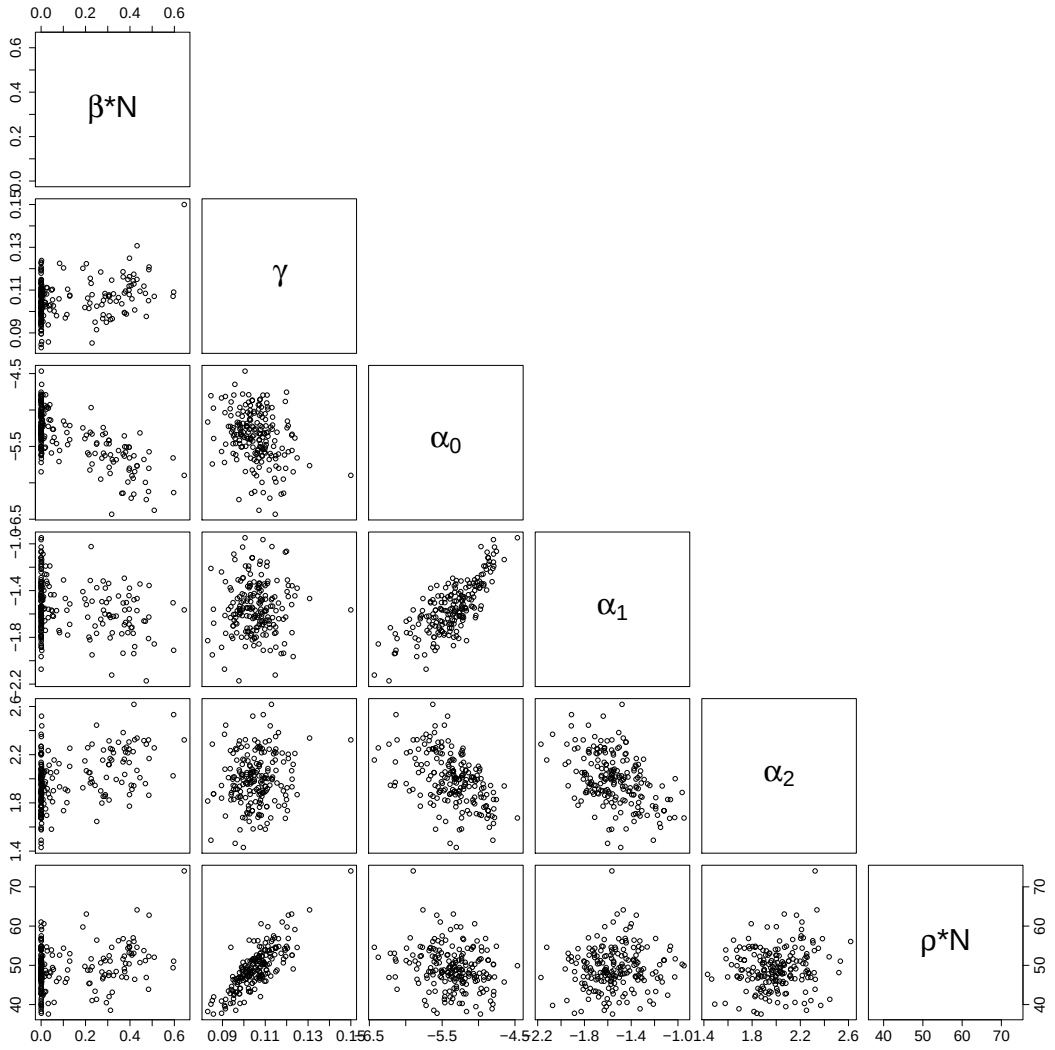

Figure 3.11: Bivariate scatterplots of parameters of the SIRS model estimated using data from Mathbaria. Scatterplots are thinned to display only 200 samples, so only every 200th sample from the posterior distribution is plotted.

Coefficient	Starting value set 1	Starting value set 2	Starting value set 3
$\beta \times N$	0.6	0.00006	0.00034
γ	0.11	0.08	0.12
α_0	-7.11	-5	-3
α_1	0	2	1
α_2	0	2	-1
$\rho \times N$	60	40	100

Table 3.2: Initial values used for separate runs of the PMMH algorithm on the data from Mathbaria.

	Starting value set 1		Starting value set 2		Starting value set 3	
Coefficient	Estimate	95% CIs	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0046	(0 , 0.5178)	0.0019	(0 , 0.4963)	0.0279	(0 , 0.5346)
γ	0.1	(0.09 , 0.12)	0.11	(0.09 , 0.12)	0.11	(0.09 , 0.12)
$(\beta \times N)/\gamma$	0.04	(0 , 4.65)	0.02	(0 , 4.5)	0.27	(0 , 4.76)
α_0	-5.35	(-6.12 , -4.75)	-5.31	(-6.07 , -4.71)	-5.38	(-6.14 , -4.75)
α_1	-1.56	(-1.98 , -1.12)	-1.55	(-1.97 , -1.11)	-1.56	(-1.98 , -1.13)
α_2	1.98	(1.58 , 2.44)	1.97	(1.58 , 2.41)	2	(1.59 , 2.46)
$\rho \times N$	48.8	(39.6 , 60.31)	48.66	(39.43 , 59.87)	48.79	(39.84, 60.45)
	N=10000, $1/\mu = 2$ years $\phi_S/N = 0.2$, $\phi_I/N = 0.02$		N=5000, $1/\mu = 3$ years $\phi_S/N = 0.2$, $\phi_I/N = 0.02$		N=50000, $1/\mu = 3$ years $\phi_S/N = 0.2$, $\phi_I/N = 0.02$	
Coefficient	Estimate	95% CIs	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0051	(0 , 0.5185)	0.0044	(0 , 0.5232)	0.0025	(0 , 0.3559)
γ	0.11	(0.09 , 0.13)	0.1	(0.09 , 0.12)	0.11	(0.09 , 0.13)
$(\beta \times N)/\gamma$	0.05	(0 , 4.56)	0.04	(0 , 4.77)	0.02	(0 , 3.25)
α_0	-5.25	(-6.47 , -4.72)	-5.53	(-6.25 , -5.01)	-5.31	(-5.78 , -4.96)
α_1	-1.5	(-2.01 , -1.12)	-1.71	(-2.08 , -1.35)	-1.53	(-1.83 , -1.26)
α_2	2.03	(1.63 , 2.72)	2	(1.63 , 2.43)	2.02	(1.75 , 2.32)
$\rho \times N$	36.36	(29.82, 44.8)	48.16	(39.3 , 59.32)	50.01	(41.2 , 60.99)
	N=10000, $1/\mu = 3$ years $\phi_S/N = 0.4$, $\phi_I/N = 0.04$		N=10000, $1/\mu = 3$ years $\phi_S/N = 0.1$, $\phi_I/N = 0.01$		N=10000, $1/\mu = 3$ years $\phi_S/N = 0.4$, $\phi_I/N = 0.01$	
Coefficient	Estimate	95% CIs	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0416	(0 , 0.2514)	0.677	(0.0045, 0.9056)	0.001	(0 , 0.2388)
γ	0.11	(0.09 , 0.13)	0.11	(0.09 , 0.13)	0.11	(0.09 , 0.13)
$(\beta \times N)/\gamma$	0.39	(0 , 2.24)	6.33	(0.04 , 7.91)	0.01	(0 , 2.16)
α_0	-5.48	(-6.2 , -4.8)	-6.05	(-6.74 , -5.02)	-5.41	(-6.09 , -4.78)
α_1	-1.64	(-2.14 , -1.08)	-1.69	(-2.16 , -1.27)	-1.63	(-2.11 , -1.12)
α_2	1.89	(1.43 , 2.31)	2.49	(1.85 , 3.09)	1.91	(1.47 , 2.3)
$\rho \times N$	41.63	(33.48, 50.95)	55.74	(45.24 , 68.52)	41.76	(34.13, 51.43)

Table 3.3: Convergence diagnostics and sensitivity analysis: Posterior medians and 95% equitailed credible intervals (CIs) under different initial values and assumptions for the parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh. The PMMH algorithm is run from different initial values using $N = 10000$, $1/\mu = 3$ years, $\phi_S/N = 0.2$, and $\phi_I/N = 0.02$, and also run using different values for the population size, N, the loss of immunity rate, μ, and the means of the Poisson initial distributions, ϕ_S and ϕ_I.

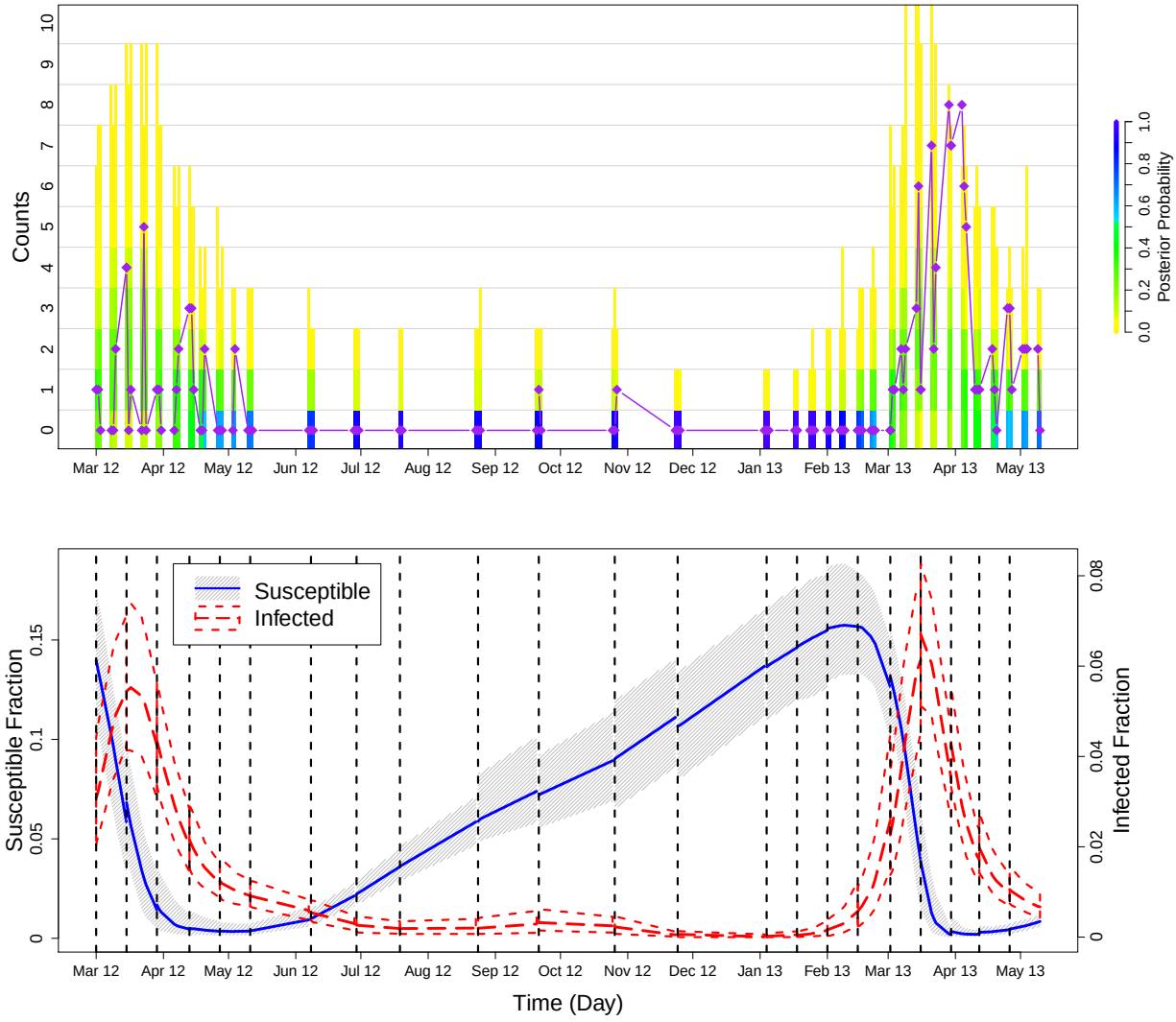

Figure 3.12: Summary of prediction results for the second to last and last epidemic peaks in the Bangladesh data. We again run PMMH algorithms on training sets of the data, which are cut off at each of the dashed black lines in the bottom plot, and future cases are predicted until the next cut off. The top plot compares the posterior probability of the predicted counts to the test data (purple diamonds and line), and the bottom plot shows how the trajectory of the predicted hidden states changes over the course of the epidemic. See the caption of Figure 3.7 for more details.

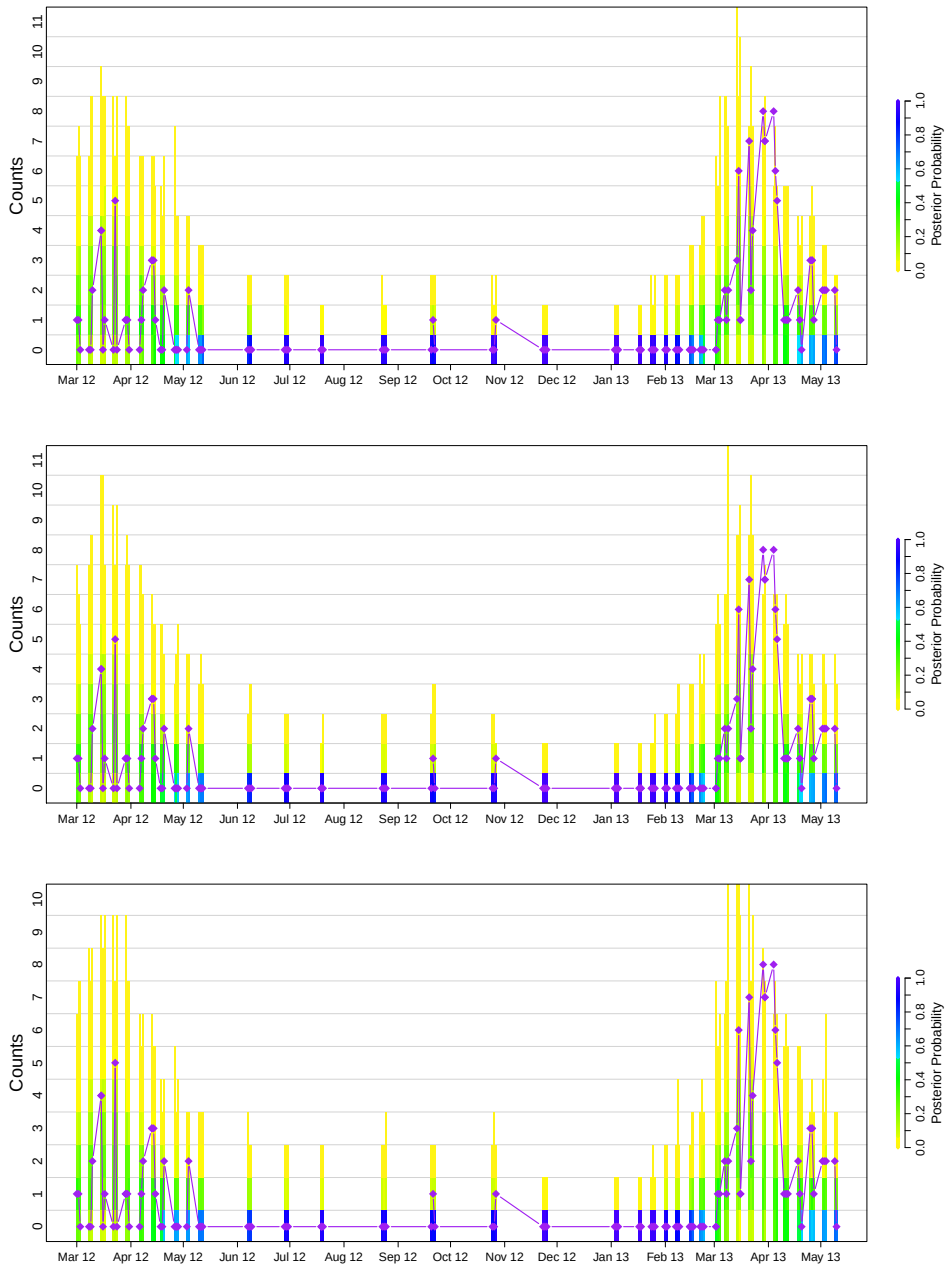

Figure 3.13: Distributions of predicted reported cases under models assuming a covariate lag of 14 days (top), 18 days (middle), and 21 days (bottom). The posterior probability of the predicted counts is compared to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The distributions are similar, regardless of lag choice.

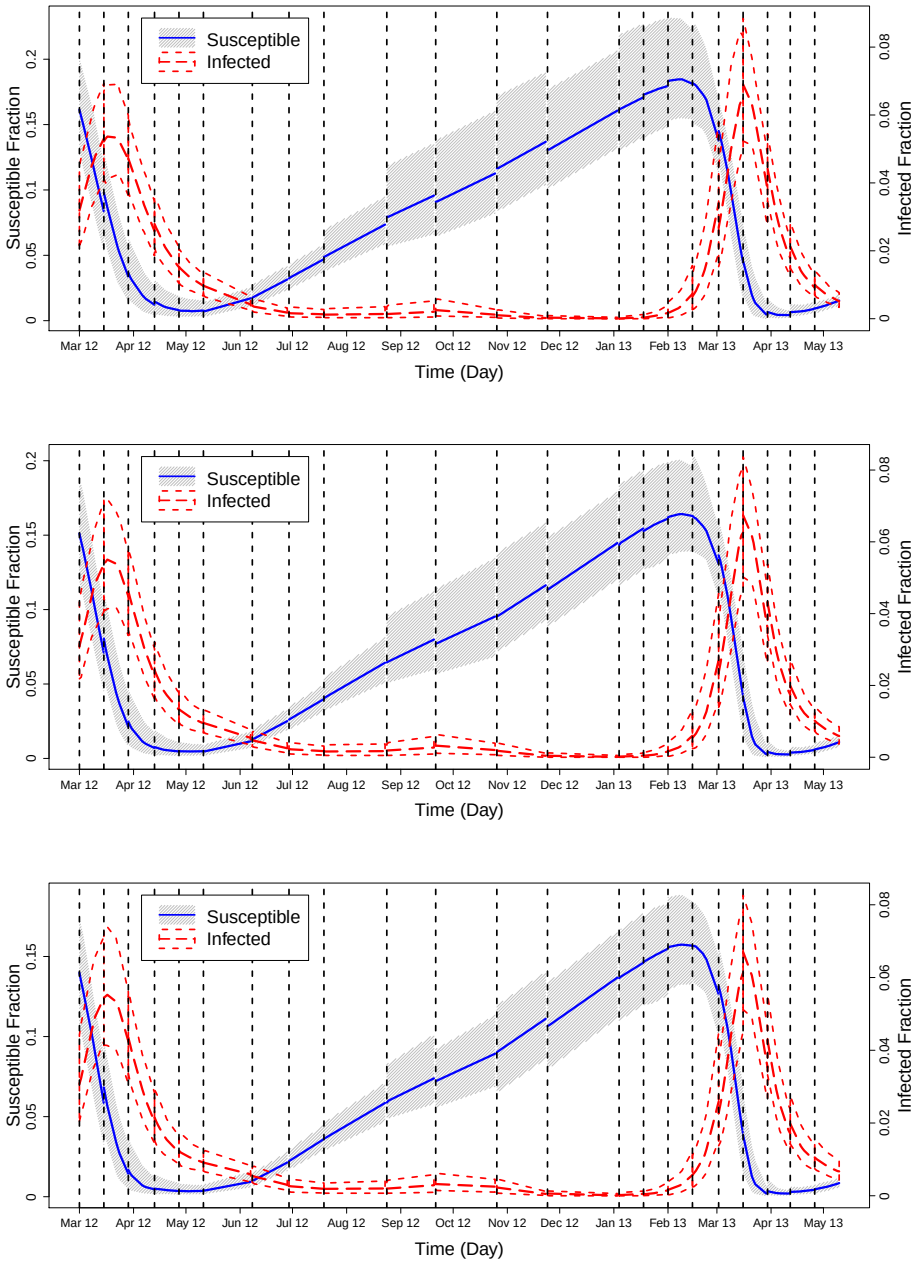

Figure 3.14: Predictive distributions of the hidden states, under models assuming a covariate lag of 14 days (top), 18 days (middle), and 21 days (bottom). The gray area and the solid line denote the 95% quantiles and median, respectively, of the predictive distributions for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median, respectively, of the predictive distributions for the fraction of infected individuals. Differences between the distributions are difficult to distinguish.

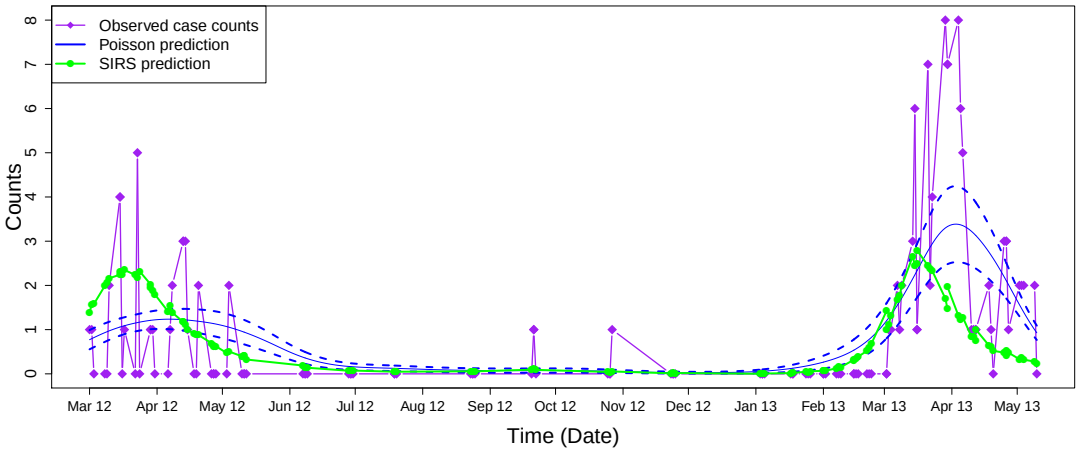

Figure 3.15: Comparison of predicted means for number of reported cases. The solid blue lines and the dashed blue lines denote the predicted means and 95% intervals, respectively, under the quasi-Poisson model. The green dots (connected by a green line) denote the predicted means under the SIRS model. Predictions are started and stopped using identical cut-off points for the training and test data to those in Figure 3.12. Test data is denoted by the purple diamonds connected by straight lines.

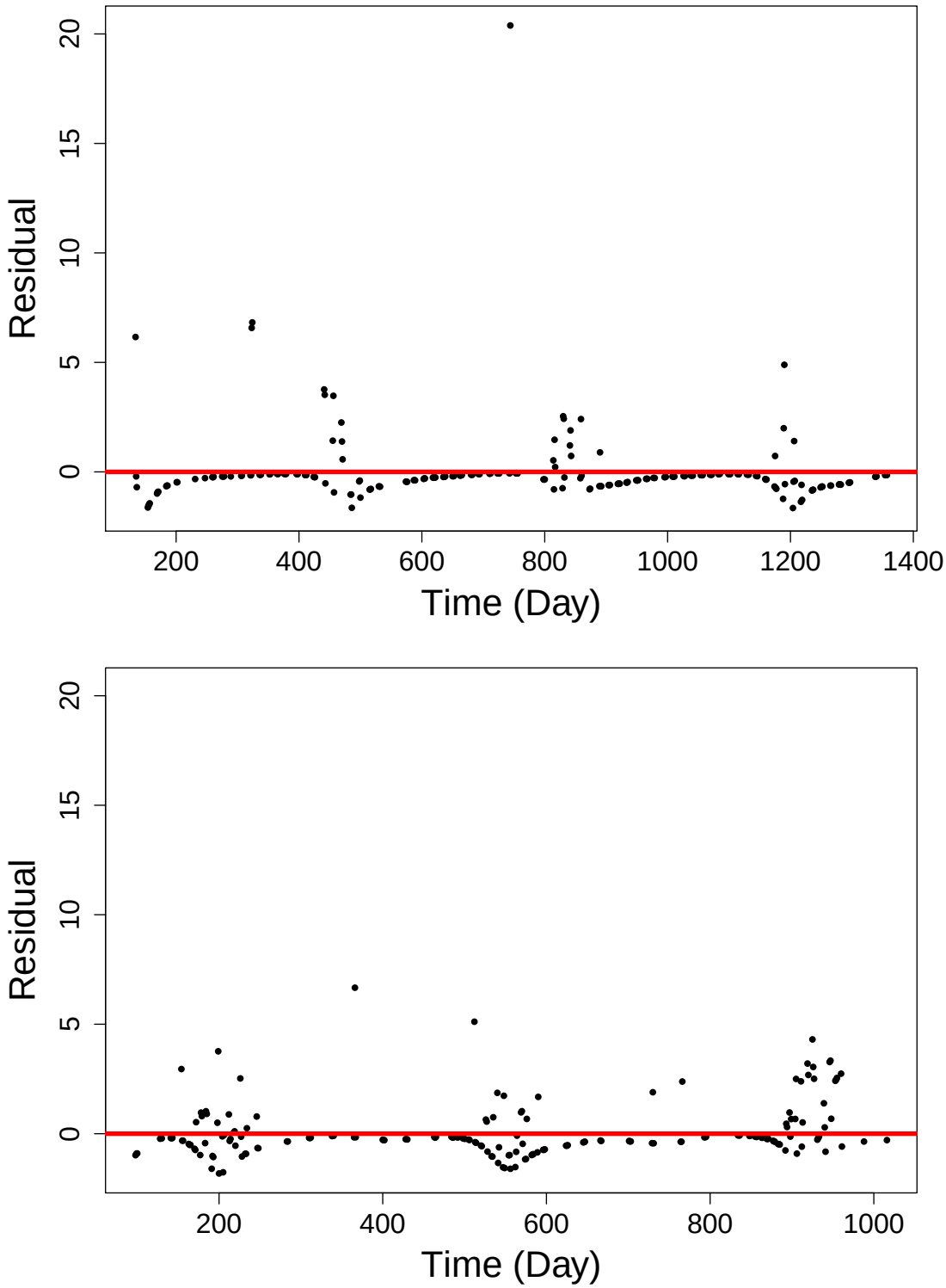

Figure 3.16: Plot of standardized residuals versus time. The top figure shows the residuals for the first three years of data collected from Bangladesh, and the bottom figure shows the residuals for the second three years of data collection. The red line is drawn through zero for reference.

Chapter 4

SIRS ANALYSIS OF CHOLERA OUTBREAKS IN OTHER STUDY AREAS IN BANGLADESH

We seek to develop a predictive model for cholera outbreaks that could be extended to multiple geographical areas in Bangladesh. To do this, we first need to test the generalizability of our methods and results. Using the model and methods described in Chapter 3, we analyze cholera incidence data and ecological data collected from multiple thanas in rural Bangladesh over the past sixteen years. Figure 4.1 shows a map of the study areas. There have been three phases of data collection so far, each lasting approximately three years and being separated by gaps of a few years; the current collection phase is ongoing. The first phase spans March 1997 to December 2001 and included data collected from four study sites: Bakerganj, Chaugachha, Chhatak, and Matlab. Details of surveillance in these study areas during this phase are described by Sack et al. [2003]. The second phase of data collection lasted from 2004 to 2008 and was conducted in Mathbaria and Bakerganj. The third phase refocused on Chhatak and Mathbaria, started in 2010, and will continue until 2015. During these time periods, cholera incidence data were collected over a period of three days approximately every two weeks. Environmental data were also collected approximately every two weeks from four to six water bodies. The same water bodies were not always used for the entire study period; some sample sites were dropped and others were picked up, but sampling from four to six sites was maintained. Water bodies include rivers, lakes, and ponds. Some water bodies are control sites, meaning they are well regulated, used only for drinking water, and kept separate from salt water. Thus there is limited contamination of these control water bodies by individuals infected with cholera. Over the phases, different environmental variables have been sampled; Table 4.1 shows all covariates that were sampled during the three phases and when they were sampled.

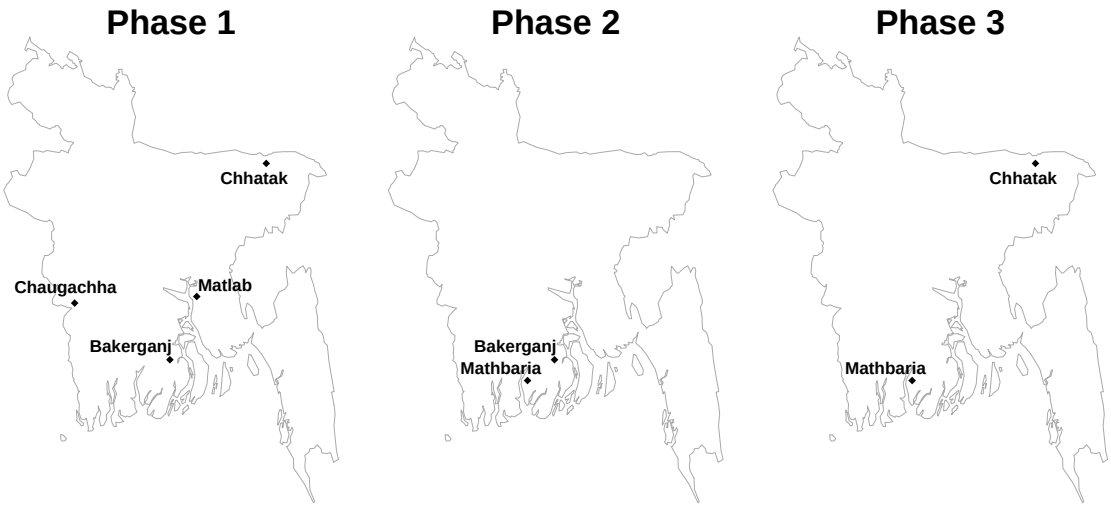

Figure 4.1: Map of Bangladesh study sites. There have been three phases of data collection so far, focusing on different geographical areas of Bangladesh. Each map denotes the study areas observed in that phase.

Huq et al. [2005] used Poisson regression to study the association between lagged predictors from a particular water body to cholera cases in the thana containing that water body. Denoting Y_t as the number of cholera cases at time t and $X_{i,t-\tau_i}$ as the ith lagged predictor at time $t-\tau_i$, they selected τ_i such that $X_{i,t-\tau_i}$ had the strongest correlation with Y_t, for all possible τ_i. They then used a Poisson regression for k predictors, so that

$$\ln \mathrm{E}(Y_t|X) = \beta_0 + \beta_1 X_{1,t-\tau_1} + \beta_2 X_{2,t-\tau_2} + ... + \beta_k X_{k,t-\tau_k},$$

and used stepwise regression for variable selection. This resulted in different lags and different significant covariates across multiple water bodies and thanas. Environmental covariates included water conductivity, water temperature, water depth, pH, air temperature, salinity, cholera toxin probe-positive count, dissolved oxygen, total rainfall, copepod counts, cyanobacteria, and fecal coliform counts.

We use a Bayesian framework to estimate the parameters of our hidden SIRS model using data from the Phase 1 study sites Bakerganj, Chaugachha, Chhatak, and Matlab, as we did with data from Mathbaria in Chapter 3. We again look at the environmental

	Phase		
Covariates	1	2	3
Water conductivity μS/cm	*	*	*
Combined Copepod counts ($\log_{10}$ count/ml)	*		
Adult Copepod counts (count/L)		*	*
Juvenile Copepod counts (count/L)		*	*
ctx probe-positive ($\log_{10}$ count/ml)	*		
Water temperature (°C)	*	*	*
Air temperature (°C)	*	*	*
Water depth (m)	*	*	*
Water pH	*	*	*
Cyanobacteria ($\log_{10}$ count/ml)	*		
Dissolved O^2 (mg/liter)	*	*	*
Fecal coliforms ($\log_{10}$ CFU/ml)	*		
Water salinity (ppt)	*	*	*
Total rainfall (mm/2 weeks)	*		
Total dissolved solids (mg/L)		*	*
Turbidity			*

Table 4.1: Environmental variables sampled during the three phases of data collection in Bangladesh. An asterisk indicates that sampling occurred for that variable in that phase.

covariates water depth and water temperature and compare covariate estimates from these geographically varied areas in Bangladesh.

Future work will focus on combining information across the study sites to develop a predictive model that could work in any area of Bangladesh, or be extended to other places. To lay the foundation for this, we first focus our attention on a couple of covariates and look at estimates from each study area separately to give us an idea of how generalizable our results are. We want to see if predictive covariates have similar relationships with cholera outbreaks across study sites.

Figures 4.2 and 4.3 show water depth and water temperature measurements for the Phase 1 data. We fit a cubic spline to these covariates to get a smooth summary of covariate behaviors; we want to combine information from all study sites within a thana, and the average over the sites is too noisy. Because we do not know which cholera cases were exposed to which pond, we are just interested in a surrogate for the underlying covariate behavior in the environment rather than specific measurements from the water bodies.

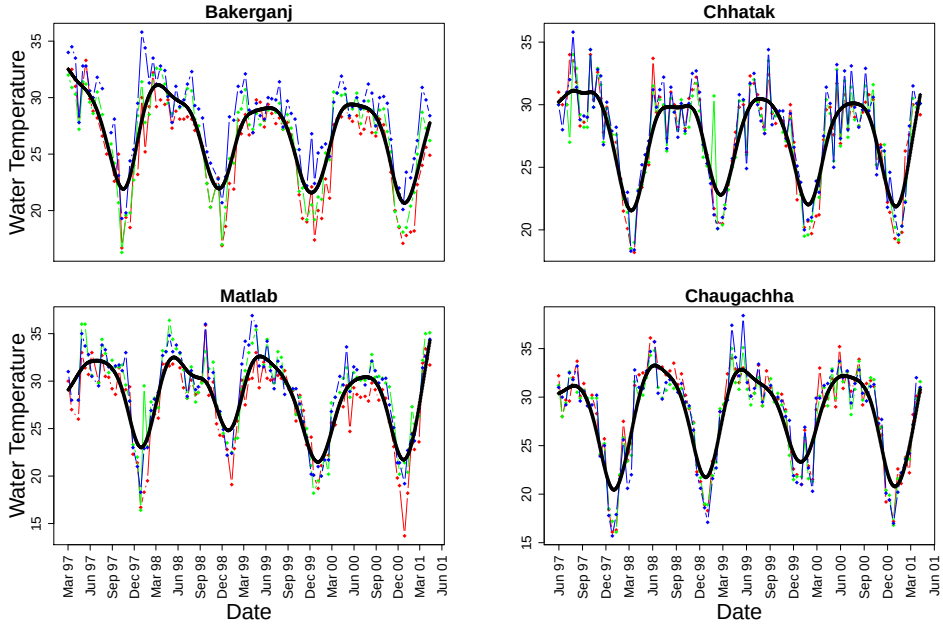

Figure 4.2: Water temperature measurements in the four thanas studied in the first phase of data collection. Each color denotes measurements from a specific study site, and the black line denotes a cubic smoother through the measurements.

To compare estimates with the results of Chapter 3, we again consider a covariate lag $\kappa = 21$ days. Using daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$, the environmental force of infection

$$\alpha(t) = \alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 C_{WD}(i-\kappa) + \alpha_2 C_{WT}(i-\kappa)\right] \text{ for } t \in A_i \text{ and } t \geq \kappa.$$

Again $C_{WT}(i) = (WT(i) - \overline{WT})/s_{WT}$ and $C_{WD}(i) = (WD(i) - \overline{WD})/s_{WD}$, where $\overline{X}$ is the mean of the measurements for all i and s_X is the sample standard deviation. Figure 4.4 shows cholera case counts and the smoothed, lagged covariate values for the four study sites. We use only data collected between March 1997 and March 2001 because that is when data are available for all four study areas.

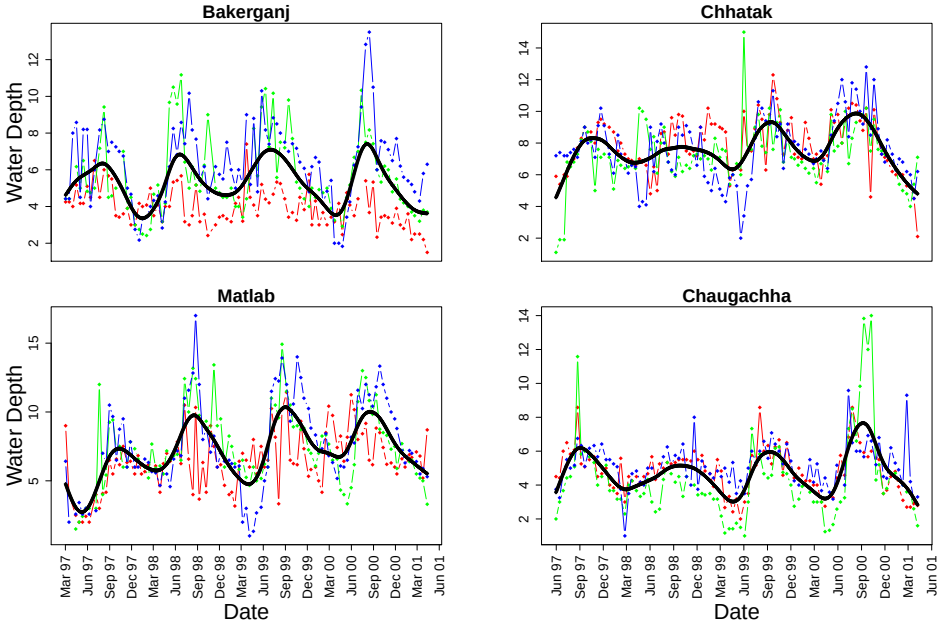

Figure 4.3: Water depth measurements in the four thanas studied in the first phase of data collection. Each color denotes measurements from a specific study site, and the black line denotes a cubic smoother through the measurements.

4.1 Comprehensive Bayesian analysis of Phase 1 data

For Bayesian analysis on all four study areas, we use the same prior distributions, number of iterations and particles, thinning, and parameter assumptions. We assume diffuse Normal prior distributions on all of our transformed parameter values, using transformations for parameters that are constrained to be greater than zero or between zero and one. Prior distributions for $\log(\beta)$ and α_0 are centered at $\log(1.25 \times 10^{-7})$ and -8, respectively, with standard deviations of 5. Priors for the time-varying environmental covariates α_1 and α_2 are centered at 0 and have standard deviations of 5. The prior for $\log(\gamma)$ is constructed such that the average infectious period for cholera, $1/\gamma$, is between 8 and 12 days; thus the prior mean for $\log(\gamma)$ is $\log(0.1)$ and the prior standard deviation is 0.09. We also incorporate information into the prior on $\text{logit}(\rho)$, since we know that the number of symptomatic cholera cases should be low and even fewer will be reported [Sack et al., 2003]. The prior

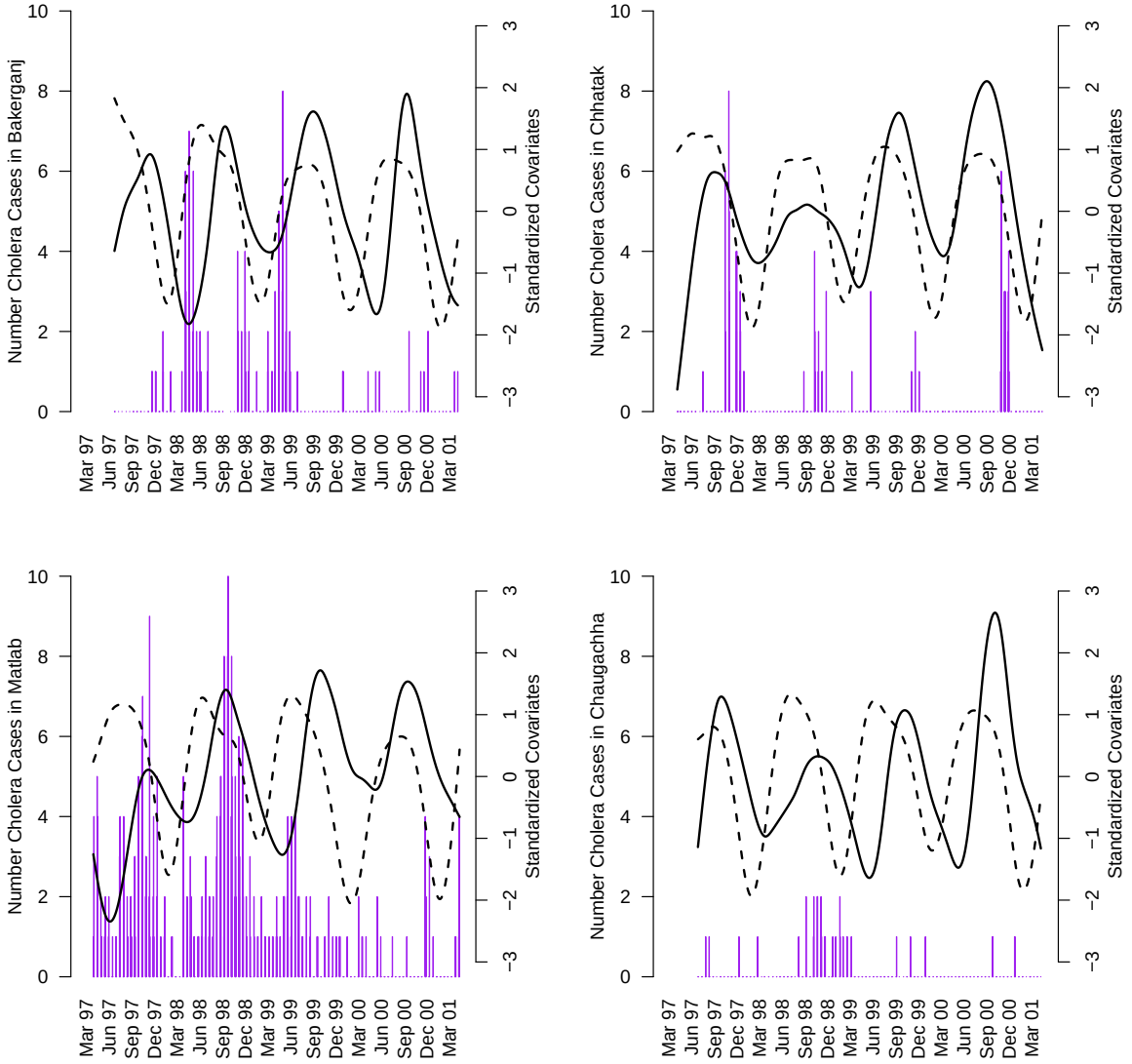

Figure 4.4: Barplots of cholera case counts in the four thanas studied in the first phase of data collection and the standardized covariate measurements over time.

mean for logit(ρ) is logit(8×10^{-4}) and the prior standard deviation is 2.

We set the population size N=10000, and we set $\phi_S = 2000$ and $\phi_I = 200$. The PMMH algorithm starts with a burn-in run of 30000 iterations, a secondary run of 20000 iterations, and a final run of 400000 iterations. To thin the chains, we save only every 10th iteration.

We use $K = 100$ particles in the SMC algorithm.

Table 4.2 shows effective sample sizes of all model parameters. It is clear that the PMMH algorithm applied to the data from Bakerganj and Matlab did not have good mixing; we discuss this in the following sections. Posterior medians and 95% credible intervals for the parameters of the SIRS model estimated using data from Chhatak and Chaugachha are given in Table 4.3. For comparison, we include results from the Mathbaria analysis presented in Chapter 3. Overall, posterior medians and credible intervals vary widely, except for the parameter γ which has a very strong prior distribution. The posterior for the reporting rate ρ varies, which makes sense as this parameter should account for different population sizes in the four thanas and reflect possible model misspecifications. We describe the details of the results by study area in the following sections, focusing on the parameters relating to the environmental force of infection $\alpha(t)$.

Table 4.2: Effective sample sizes for the parameters of the SIRS model estimated using clinical and environmental data sampled from four different sample sites in Bangladesh.

	Effective sample size					
	β	γ	α_0	α_1	α_2	ρ
Chaugachha	209	1720	21	188	101	250
Chhatak	545	13738	19937	18978	20092	14503
Matlab	1	262	2	2	3	3
Bakerganj	0	5878	614	184	356	2440

4.1.1 Chaugachha

Chaugachha was selected as a control area, since historically there were no cholera outbreaks in the area [Sack et al., 2003]. Summary plots of the PMMH algorithm output are in Figure 4.5. Most of the parameters have bi-modal marginal posterior distributions. This could be due to the low number of cholera cases observed in Chaugachha during the study period. The posterior credible intervals for the environmental parameters are very wide and cover zero, so water depth and water temperature do not have a significant relationship with the force of infection in Chaugachha. This makes sense; there are very few cholera cases in this

Table 4.3: Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using clinical and environmental data sampled from three study sites in Bangladesh.

	Chhatak		Chaugachha	
Coefficient	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0004	(0 , 0.0896)	0.0142	(0 , 1.2825)
γ	0.09	(0.08 , 0.11)	0.1	(0.08 , 0.12)
$(\beta \times N)/\gamma$	0.004	(0 , 0.986)	0.142	(0 , 14.495)
α_0	-0.76	(-1.58 , 0.41)	-2.79	(-17.36, -0.21)
α_1	-0.77	(-1.29 , -0.32)	1.78	(-4.38 , 4.62)
α_2	-5.74	(-7.36 , -4.59)	-2.82	(-8.9 , 2.7)
$\rho \times N$	29.21	(22.47, 37.63)	13.87	(9.1 , 25.33)

	Mathbaria	
Coefficient	Estimate	95% CIs
$\beta \times N$	0.0046	(0 , 0.5178)
γ	0.1	(0.09 , 0.12)
$(\beta \times N)/\gamma$	0.044	(0 , 4.648)
α_0	-5.35	(-6.12 , -4.75)
α_1	-1.56	(-1.98 , -1.12)
α_2	1.98	(1.58 , 2.44)
$\rho \times N$	48.8	(39.6 , 60.31)

short time series, so there is low power to detect significant relationships with the force of infection.

4.1.2 Chhatak

Using the data from Chhatak, both water depth and water temperature have a significant relationship with the force of infection since the credible intervals for α_1 and α_2 do not include zero. Decreasing water depth increases the force of infection, and decreasing water temperature increases the force of infection. This relationship for water temperature is opposite of what we saw using the data from Mathbaria, and opposite of what we would expect biologically.

4.1.3 Matlab

From the trace plots in the top half of Figure 4.7 and the effective sample sizes in Table 4.2, it is clear we had problems with SIRS parameter estimation using the data from Matlab, in particular with the mixing in the PMMH algorithm. In the original final run, the algorithm found an area of high posterior likelihood, and once the PMMH algorithm started sampling from that area of the posterior distribution accepting new proposals became very difficult.

Using a modified final run of the PMMH algorithm, we restart the Markov chain in this new high likelihood area and retune the proposal distribution. Using the first 100000 iterations of the original final run, we calculate the approximate posterior covariance of the parameters and use this to construct the covariance of the multivariate normal proposal distribution in the modified final run of 40000 iterations of the PMMH algorithm. We increase the number of particles to $K = 1000$ to decrease the variance in the estimate of the likelihood. Trace plots for the parameters are shown in the bottom half of Figure 4.7. Mixing improves slightly, but effective sample sizes for the parameters are still too small to draw any conclusions, as seen in Table 4.4.

4.1.4 Bakerganj

We also had problems with mixing in the PMMH algorithm using the data from Bakerganj, as seen in the trace plots in the top half of Figure 4.8 and the effective sample sizes in Table 4.2. In particular, the mixing for β is very poor; it is clear this chain has not converged.

To improve mixing, we again use the original final run to calculate the approximate posterior covariance of the parameters and use this to construct the covariance of the multivariate normal proposal distribution in a modified final run of 90000 iterations of the PMMH algorithm. To thin the chains, we save only every 10th iteration. We increase the number of particles to $K = 500$ in the SMC algorithm. We restart the Markov chain where the parameters were mixing well in the original final run, but start β at a lower value since it is clear from the trace plot in top half of Figure 4.8 that β is moving towards lower values. Trace plots for the parameters generated by the modified final run are shown in the bottom half of Figure 4.8. It is clear β still is not mixing well, and now the effective sample sizes

for γ and ρ are also very small, as seen in Table 4.4.

Typically, Bakerganj has a large cholera outbreak in the spring (pre-monsoon) and a smaller peak in the late fall (post-monsoon), as seen in Figure 4.4. These peaks are missing after September 1999, which may be affecting estimation. There also appear to be less pronounced peaks in Matlab after September 1999. It is clear this is a bad model choice for the data looking at the plot of cases and covariates. We suspect that this model misspecification manifests itself in poor mixing of the PMMH MCMC.

Table 4.4: Effective sample sizes for the parameters of the SIRS model generated from modified final PMMH runs using data from Matlab and Bakerganj

	Effective sample size					
	β	γ	α_0	α_1	α_2	ρ
Matlab	0	6	24	13	4	10
Bakerganj	0	71	445	1255	856	1

4.2 Discussion

Overall, the posterior estimates for the parameters of the SIRS model vary widely between thanas. The two covariate model is not easily generalizable as there seems to be a large amount of heterogeneity even within Bangladesh in how these covariates are related to cholera outbreaks. This result is not entirely surprising, as we are not sure how these covariates operate across thanas. Water depth decreasing in the southern part of Bangladesh affects *V. cholerae* growth differently than a similar decrease in water depth in the northern thanas [Akanda et al., 2009, 2011, Jutla et al., 2013]. Akanda et al. [2011] found that, in the Bengal Delta region, both the tidal intrusion of *V. cholerae* contaminated water pre-monsoon and flooding post-monsoon are related to cholera outbreaks, causing biannual peaks in this region. These dual peaks are not seen in the northern thanas. We will need to keep this in mind when trying to extend our SIRS model to do prediction of cholera outbreaks in other areas.

Further fine tuning is needed for the analysis of the data from Matlab and Bakerganj. It is clear the current techniques are not working; these problems may be solved with a better

MCMC proposal distribution for the SIRS model parameters or by a more efficient way to initialize the PMMH algorithm. Incorporating addition data, such as the data collected in Bakerganj during Phase 2, might also help. We will test this in future analyses.

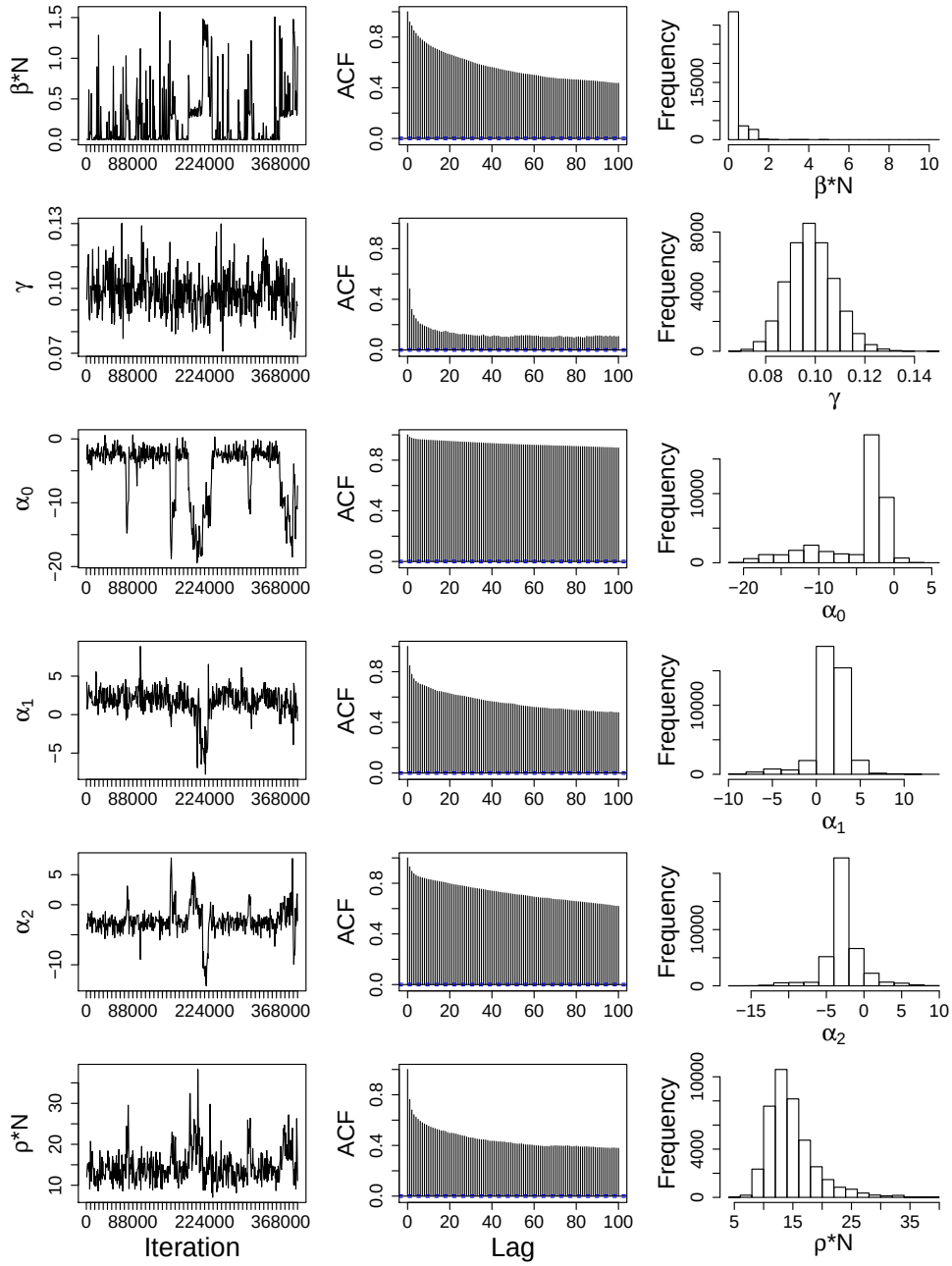

Figure 4.5: Trace plots, auto-correlation plots, and histograms for the parameters of the SIRS model from the final run (400000 iterations) of the PMMH algorithm using data from Chaugachha. ACF plots and histograms are thinned to 40000 iterations and trace plots are thinned to display only 500 iterations. The bi-modal posterior distributions, especially seen in the environmental covariates, could be a product of the lack of cholera outbreaks in Chaugachha.

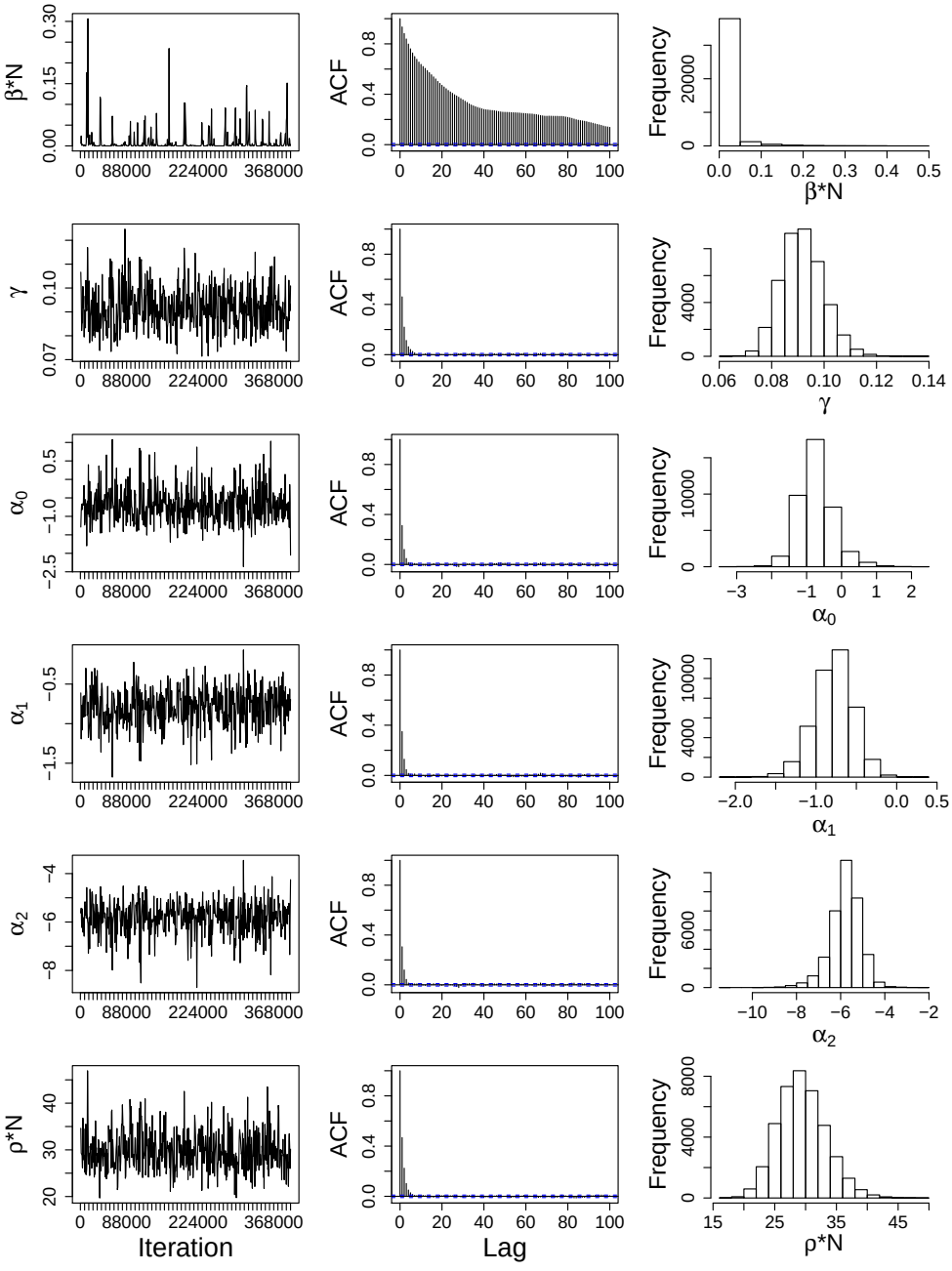

Figure 4.6: Trace plots, auto-correlation plots, and histograms for the parameters of the SIRS model from the final run (400000 iterations) of the PMMH algorithm using data from Chhatak. ACF plots and histograms are thinned to 40000 iterations and trace plots are thinned to display only 500 iterations.

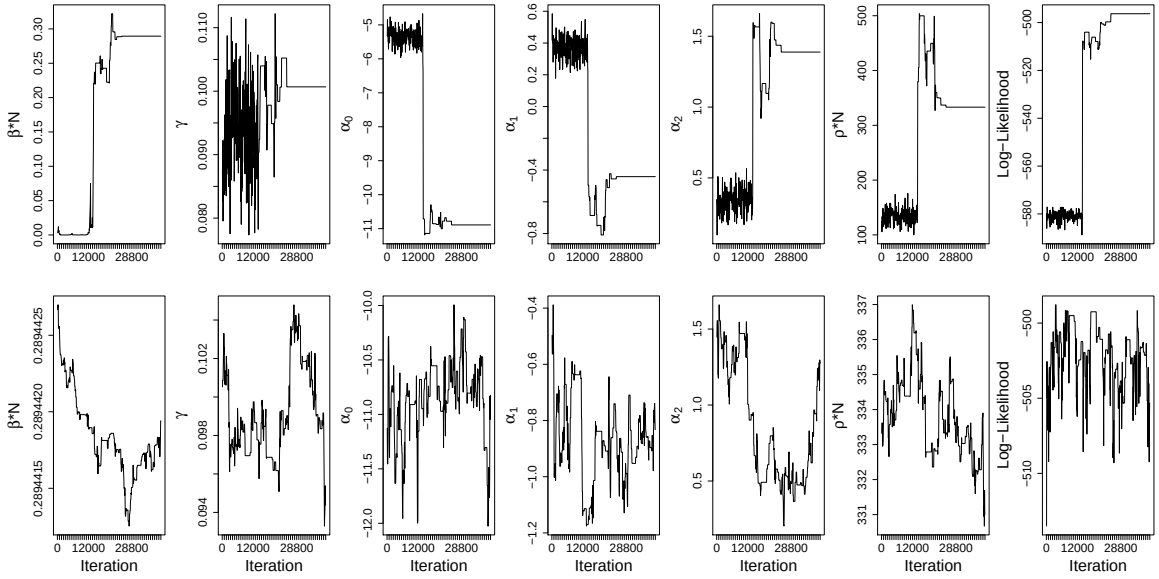

Figure 4.7: Trace plots for the parameters of the SIRS model from the original final run (top row) and the modified final run (bottom row) of the PMMH algorithm using data from Matlab. Plots are thinned to display only 500 iterations.

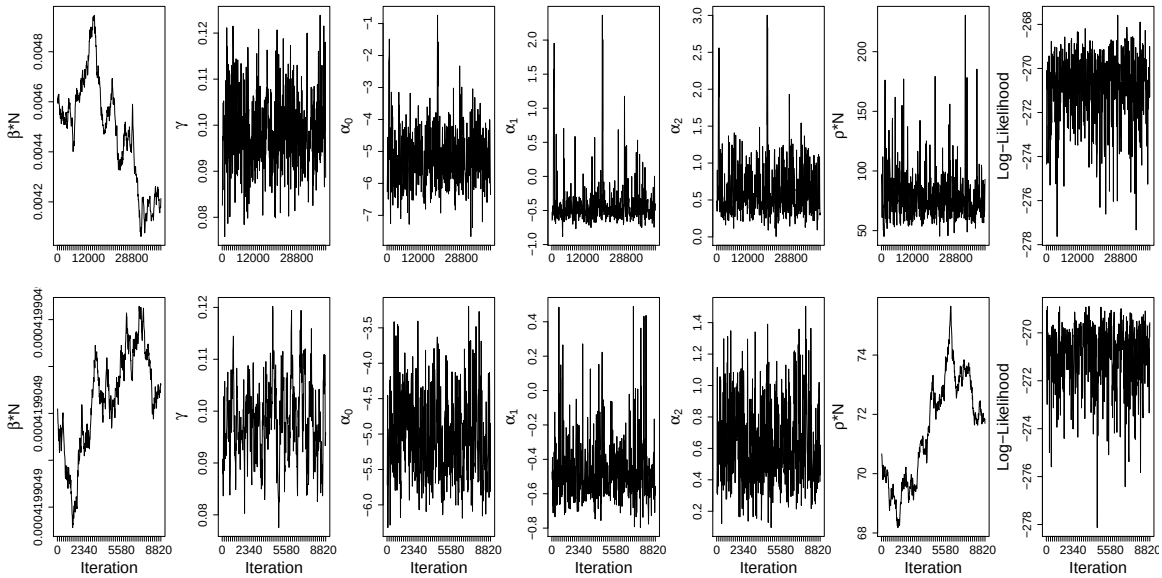

Figure 4.8: Trace plots for the parameters of the SIRS model from the original final run (top row) and the modified final run (bottom row) of the PMMH algorithm using data from Bakerganj. Plots are thinned to display only 500 iterations. Problems with mixing in the PMMH algorithm can be seen in the trace plot for β.

Chapter 5

FURTHER MATHBARIA ANALYSIS

Under the assumption of the SIRS model, we extend our analysis of the Mathbaria data by adding covariates and lags, testing how many covariates can be estimated using our current Bayesian framework. The environmental covariates of interest are similar to the covariates analyzed by Huq et al. [2005]: water temperature, water depth, water conductivity, pH, air temperature, salinity, total dissolved solids (TDS), and copepod counts. As shown in Table 4.1, cyanobacteria and cholera toxin probe-positive counts were not collected in Phases 2 or 3, so they are excluded as covariates for the Mathbaria analysis. Initial analysis of the covariates shows very strong correlation between conductivity, TDS, and salinity. Also, conductivity is a function of pH and salinity. Thus, we start with just water temperature, water depth, conductivity, adult copepod counts, and juvenile (nauplii) copepod counts. Counts for both adult and juvenile copepods are included as covariates, as *V. cholerae* has been shown to colonize copepods [Huq et al., 1983]. Each covariate is included at three different lags: 14, 18, and 21 days.

Again we use covariates and cholera incidence data from Mathbaria, Bangladesh collected during Phases 2 (April 2004 to September 2007) and 3 (October 2010 to January 2013). Cholera incidence data and environmental data were collected approximately every two weeks, as described in previous chapters. We fit a cubic spline to get a smooth summary of our covariates, measured over six water bodies.

5.1 The Model

We consider an environmental force of infection which incorporates J covariates at L different lags. To account for lagged covariate effects, let κ_u denote the length of the lag, for lags $u \in (1, \ldots, L)$. We assume daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n-1\}$ and define the environmental force of infection $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$ and $t \geq \max_u(\kappa_u)$,

where $\alpha_{A_i} = \exp\left[\alpha_0 + \sum_{v=1}^{J}\sum_{u=1}^{L}\alpha_{L(v-1)+u}C_{v,\kappa_u}(i-\kappa_u)\right]$. Thus, we want to estimate $JL+1$ parameters relating to the environmental force of infection, $(\alpha_0, \alpha_1, \ldots, \alpha_{J\times L})$. For our analysis, $J = 5$ and $L = 3$. The covariates are the smoothed standardized daily values $C_{v,\cdot}(i) = [v(i) - \overline{v}]/s_v$, where $\overline{v}$ is the mean of the measurements for all i and s_v is the sample standard deviation. This is slightly different from how the covariates were standardized in Section 3.6, where Phases were standardized independently. Applying this global standardization to the Chapter 3 analysis does not change the results much, according to preliminary examination. The smoothed, standardized, 14 day lagged covariates and cholera incidence data are shown in Figure 5.1.

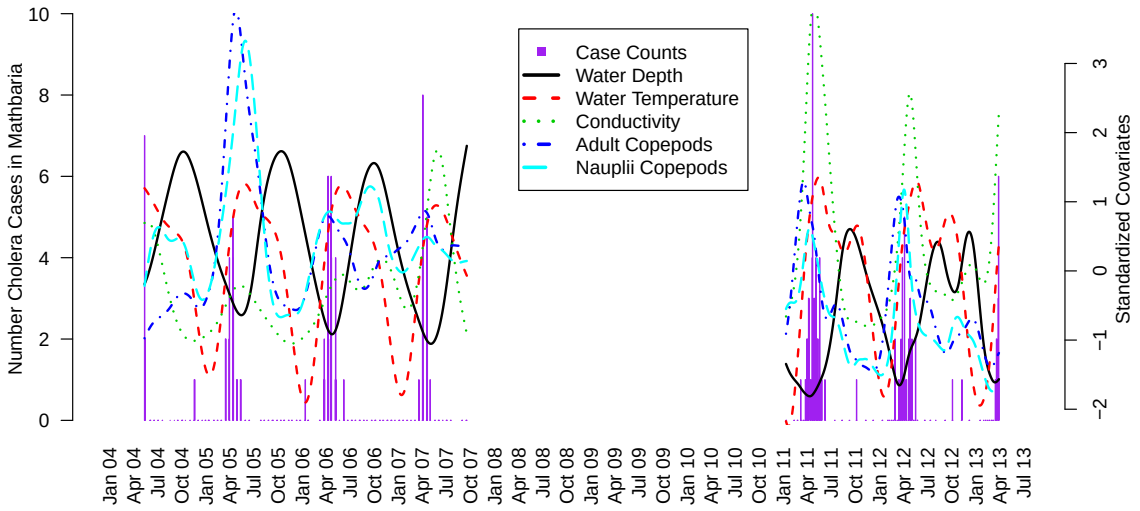

Figure 5.1: Barplot of cholera case counts in Mathbaria, Bangladesh and the standardized covariate measurements over time. The covariates are shown with a lag of two weeks. No data were collected from October 2007 through November 2010.

These 15 covariates are highly correlated, since they all vary seasonally and we use multiple lags of the same covariate in our model. Multicollinearity is an issue often encountered in regression analyses, causing overfitting and resulting in poor prediction for data outside of the training data, especially if the covariate relationships are different in the test data. We use sparsity inducing priors to address possible problems associated with multicollinearity

in our analysis.

In Chapters 3 and 4, we assume independent Normal prior distributions on α_0, α_1 and α_2 with standard deviations of 5. Here we use sparsity inducing priors to penalize covariates *a priori*, hypothesizing that, as less important covariates drop out, we will get a clearer picture of which covariate combinations are important for prediction. To see which covariates have enough signal to survive extreme prior influence, we assume increasingly small standard deviations on our prior distributions and compare estimation and prediction results. First, we assume independent Normal prior distributions on all α_k parameters for $k = 0, \ldots, J \times L$ with varying standard deviations of 5, 2, 1, and 0.5. Next, we use a non-standardized t-distribution prior with 5 degrees of freedom and varying standard deviations of 3.8, 1.55, 0.71, and 0.38. The standard deviations of the non-standardized t-distributions were selected to match the comparable 95% quantiles of the Normal prior distributions that we considered. Figure 5.2 plots the non-standardized t-distribution and Normal densities considered. The non-standardized t-distribution probability mass is still concentrated around zero, but the heavier tails of this prior allow for the possibility of large coefficient values [Gelman et al., 2008, Tipping, 2001].

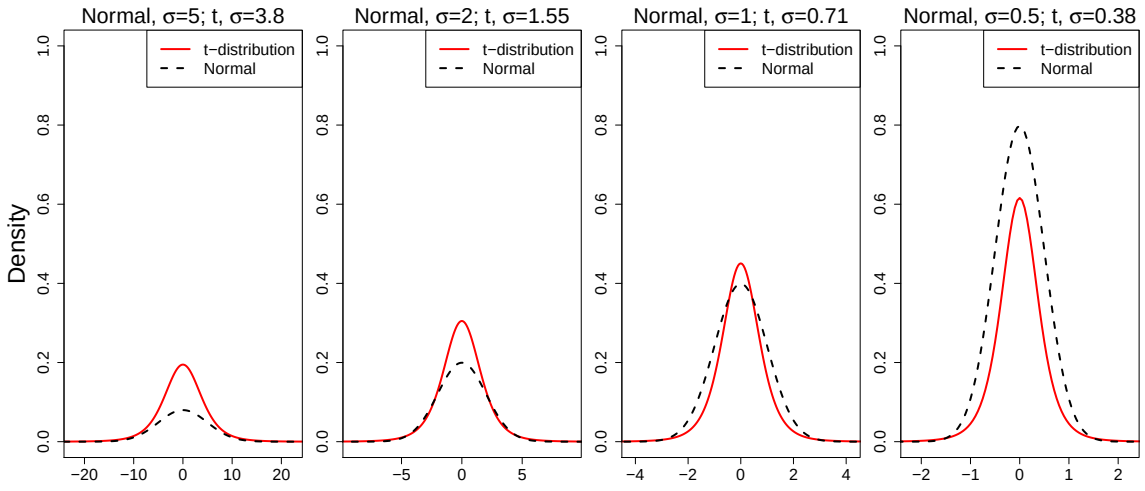

Figure 5.2: Comparison of the non-standardized t-distribution and Normal densities, under varying standard deviations.

5.2 Estimation

In the following Bayesian analyses, the prior deviation for α_0 has a mean of -8, and the priors for the k time-varying environmental covariates $\alpha_1, \ldots, \alpha_k$, are centered at 0, with varying standard deviations. We use a diffuse Normal prior distribution on the transformed parameter value $\log(\beta)$ centered at $\log(1.25 \times 10^{-7})$ and with a standard deviation of 5. We assume a Normal prior distribution for $\log(\gamma)$ with mean $\log(0.1)$ and standard deviation 0.09, so the average infectious period for cholera, $1/\gamma$, is between 8 and 12 days. The prior for the transformed parameter $\text{logit}(\rho)$ has a mean of $\text{logit}(8 \times 10^{-4})$ and standard deviation equal to 2.

As in Chapter 3, the population size is set at $N = 10000$, and ϕ_S and ϕ_I are set to 2000 and 200. The PMMH algorithm starts with a burn-in run of 30000 iterations, a secondary run of 20000 iterations, and a final run of 200000 iterations. To thin the chains, we save only every tenth iteration. We use 100 particles in the SMC algorithm.

Table 5.1 shows posterior medians and 95% credible intervals for the parameters $\beta \times N$, γ, α_0, and $\rho \times N$, generated by the final run of the PMMH algorithm under the assumption of independent Normal prior distributions for the α_k parameters with varying standard deviations. Table 5.2 shows these posterior medians and 95% credible intervals generated by the final run of the PMMH algorithm under the assumption of independent non-standardized t-distribution priors for the α_k parameters with varying standard deviations. Figure 5.3 plots the posterior medians and 95% Bayesian credible intervals for the $\alpha_k : k \in \{1, 2, \ldots, 15\}$ parameters under these different prior assumptions. From this plot, it is clear that the prior has an effect on the estimation of these parameters.

The posterior distributions of the α_k parameters show similar patterns between the comparable Normal and non-standardized t-distribution prior standard deviation assumptions. When we assume large prior standard deviations of 5 and 3.8 for the α_k parameters, we see conflicting results in the posterior distributions. For some of the covariates, such as water depth and water temperature, different lags of the same covariate have opposite relationships with the force of infection, as seen in the left plots of Figure 5.3. For example, both water depth at a lag of 14 days and water depth at a lag of 21 days have a significant

relationship with the force of infection since the credible intervals for the α_k parameters associated with water depth at these lags do not include zero. Decreasing water depth at a lag of 14 days increases the force of infection, but decreasing water depth at a lag of 21 days increases the force of infection.

These conflicting signals seem to disappear as we decrease the standard deviation of both prior distributions. For the Normal prior, when we assume the prior standard deviations of the α_k parameters are 0.5, none of the covariates have two opposing significant lags. For example, both water temperature at a lag of 14 days and water temperature at a lag of 18 days have a significant positive relationship with the force of infection.

We hypothesized that the wider tails of the non-standardized t-distribution priors would allow for improved isolation of important covariates and further elimination of conflicting signals. However, for the standard deviations of 3.8 and 1.55, the resulting significant covariates are almost identical to the comparable posteriors assuming Normal priors. The posterior distributions from the non-standardized t priors have wider credible intervals, most likely due to the wider tails in the prior distribution. These larger credible intervals wash out the significant signals seen in the two lower standard deviations under the Normal prior assumption.

From Tables 5.1 and 5.2, it is clear the posterior distributions for β and ρ vary between prior assumptions; this may be due to poor mixing in the Markov chains, as seen in Table 5.3. Mixing is better in models with smaller prior standard deviations for the α_k covariates. With wider standard deviations for these covariates, there may not be enough information in the data to estimate all 19 covariates using the current Bayesian model.

5.3 Prediction

Since our primary goal is prediction, conflicting covariate signals are not a problem if the model still predicts cholera outbreaks well. We compare prediction results from the various posterior samples. In general, using 5 covariates with 3 possible lags leads to poor prediction regardless of prior assumptions, indicating the possibility that we are overfitting our sparse time series data with a too parametrically rich model.

Figure 5.4 shows prediction results for the 2013 epidemic peak in Mathbaria under the

Table 5.1: Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh, assuming Normal prior distributions on all α_k parameters for $k = 0, \ldots, 15$ with varying standard deviations σ.

	$\Pr(\alpha_k) \sim$N(0, σ=5)		$\Pr(\alpha_k) \sim$N(0, σ=2)	
Coefficient	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0042	(0.0004, 0.141)	0.0543	(0.0023, 0.3728)
γ	0.1	(0.08 , 0.12)	0.11	(0.09 , 0.13)
$(\beta \times N)/\gamma$	0.04	(0 , 1.35)	0.48	(0.02 , 3.29)
α_0	-9.08	(-10.13 , -7.98)	-5.59	(-6.37 , -4.94)
$\rho \times N$	83.87	(51.53 , 139.44)	49.44	(40.35 , 61.25)
	$\Pr(\alpha_k) \sim$N(0, σ=1)		$\Pr(\alpha_k) \sim$N(0, σ=0.5)	
Coefficient	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.001	(0 , 0.1817)	0	(0 , 0.0029)
γ	0.11	(0.1 , 0.13)	0.12	(0.1 , 0.14)
$(\beta \times N)/\gamma$	0.01	(0 , 1.55)	0	(0 , 0.02)
α_0	-5.42	(-6.04 , -4.88)	-5.73	(-6.26 , -5.2)
$\rho \times N$	51.23	(41.72 , 62.45)	53.26	(43.16 , 65.15)

assumption of independent Normal prior distributions for the 16 α_k parameters with varying standard deviations. Figure 5.5 shows prediction results for the same epidemic peak using samples from the posterior assuming non-standardized t-distribution priors. All predictions fail to capture the severity of the increase in the fraction of infected individuals before the epidemic peak. The predictions reset most drastically at the cut-off after March 2013, where there is a large discrepancy between the predicted fraction of infected before the cut-off and the hidden states after the cut-off sampled in the PMMH algorithm. For both the Normal and non-standardized t-distribution, this discrepancy decreases as the standard deviation of the priors decreases. However, there is still a failure to predict the epidemic peak. The posterior probability of the predicted counts is not capturing the test data.

5.4 Discussion

Estimation and prediction is difficult in this framework using multiple covariates and lags. Covariate estimates highlight the problems with using multiple highly correlated variables. Prediction using the current model does not capture the increase in the fraction of infected

Table 5.2: Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh, assuming a non-standardized t-distribution on all α_k parameters for $k = 0, \ldots, 15$ with 5 degrees of freedom and varying standard deviations σ.

	σ=3.8		σ=1.55	
Coefficient	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.0001	(0 , 0.004)	0.0321	(0.0039, 0.2718)
γ	0.1	(0.08 , 0.12)	0.11	(0.09 , 0.13)
$(\beta \times N)/\gamma$	0	(0 , 0.04)	0.29	(0.03 , 2.49)
α_0	-9.02	(-10.04 , -7.01)	-5.59	(-6.4 , -4.95)
$\rho \times N$	81.35	(49.07 , 140.19)	49.56	(40.09 , 60.86)
	σ=0.71		σ=0.38	
Coefficient	Estimate	95% CIs	Estimate	95% CIs
$\beta \times N$	0.4341	(0.0058, 0.9517)	0.0037	(0.0014, 0.0791)
γ	0.12	(0.1 , 0.17)	0.11	(0.1 , 0.13)
$(\beta \times N)/\gamma$	3.76	(0.05 , 6.76)	0.03	(0.01 , 0.68)
α_0	-5.98	(-9.66 , -4.96)	-5.21	(-5.86 , -4.61)
$\rho \times N$	54.88	(43.16 , 87.81)	51.48	(41.9 , 63.44)

individuals and under-predicts the number of observed cases.

Multicollinearity of our predictive covariates could explain the poor prediction in our results. Future directions might include a hierarchical Bayesian framework, where the standard deviations of the prior distributions on the environmental covariates are estimated. However, the use of sparsity inducing priors may not be enough to fix this problem. Mixing of the Markov chain is also an issue. With so many covariates, finding a good joint proposal distribution is difficult. Using a Gibbs sampler to update parameters individually may allow for more efficient exploration of the parameter space. Once mixing is improved, we will include more covariates in the analysis of the other study sites from Bangladesh, discussed in Chapter 4.

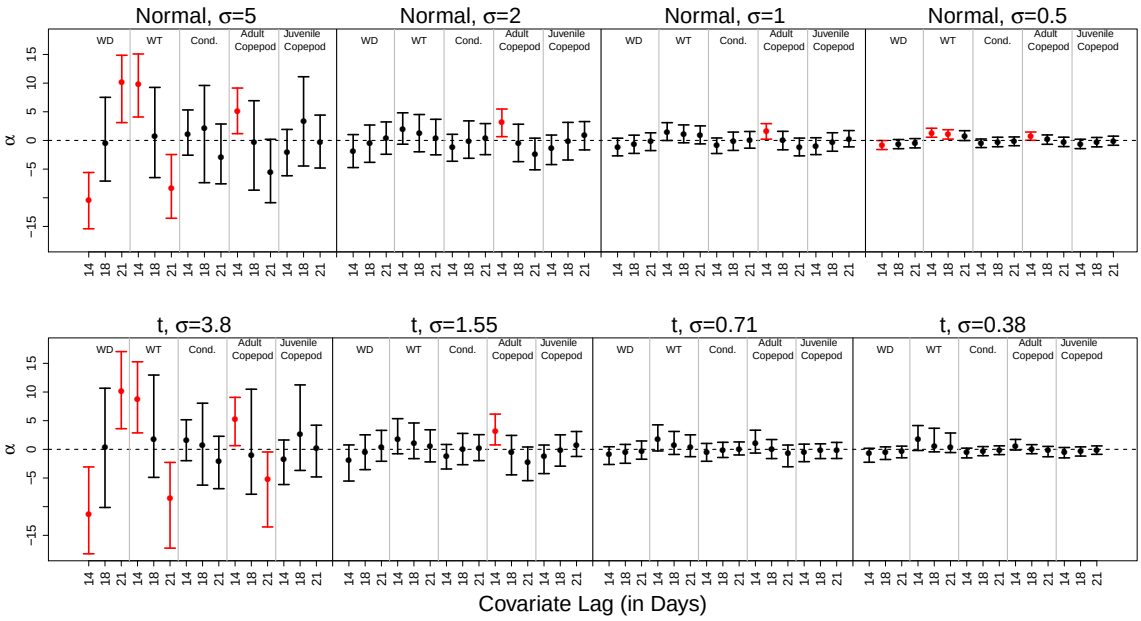

Figure 5.3: Posterior medians and 95% equitailed credible intervals for the α_k for $k = 1, \ldots, 15$ parameters of the SIRS model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh. The top row shows posterior distributions from the final run of the PMMH algorithm under the assumption of a Normal prior distribution, and the posteriors in the bottom row assume a non-standardized t-distribution with 5 degrees of freedom. Standard deviations for the non-standardized t-distribution are picked such that the 95% quantiles are similar to the Normal prior distribution in the column above it. Credible intervals not containing zero are highlighted in red.

	$\Pr(\alpha_k) \sim$N(0, σ)				$\Pr(\alpha_k) \sim$T(0, σ)			
	σ=5	σ=2	σ=1	σ=0.5	σ=3.8	σ=1.55	σ=0.71	σ=0.38
β	7	7	24	11	13	9	6	6
γ	511	499	630	635	938	955	26	409
α_0	116	555	1262	1328	90	580	5	1073
α_1	81	268	1074	845	32	215	209	663
α_2	53	367	583	1520	30	217	247	539
α_3	83	304	663	1008	53	418	243	577
α_4	81	175	1528	1646	110	418	126	611
α_5	44	299	552	1494	54	329	155	818
α_6	65	289	456	1021	80	227	222	270
α_7	132	384	1255	1117	89	505	144	956
α_8	98	142	1146	1559	82	352	362	1173
α_9	108	165	993	1623	87	591	661	615
α_{10}	184	314	982	1359	81	435	147	745
α_{11}	114	404	774	701	30	182	357	1350
α_{12}	136	371	680	1313	36	395	253	528
α_{13}	99	205	609	1094	115	237	251	2057
α_{14}	100	201	1070	974	55	401	571	1043
α_{15}	95	433	1485	2144	69	513	174	1910
ρ	147	636	1048	857	159	1397	25	520

Table 5.3: Effective sample sizes for the parameters of the SIRS model using 15 covariates in the environmental force of infection. It is clear that the chains are not mixing well, especially for the parameter β. Mixing improves for prior distributions with smaller standard deviations.

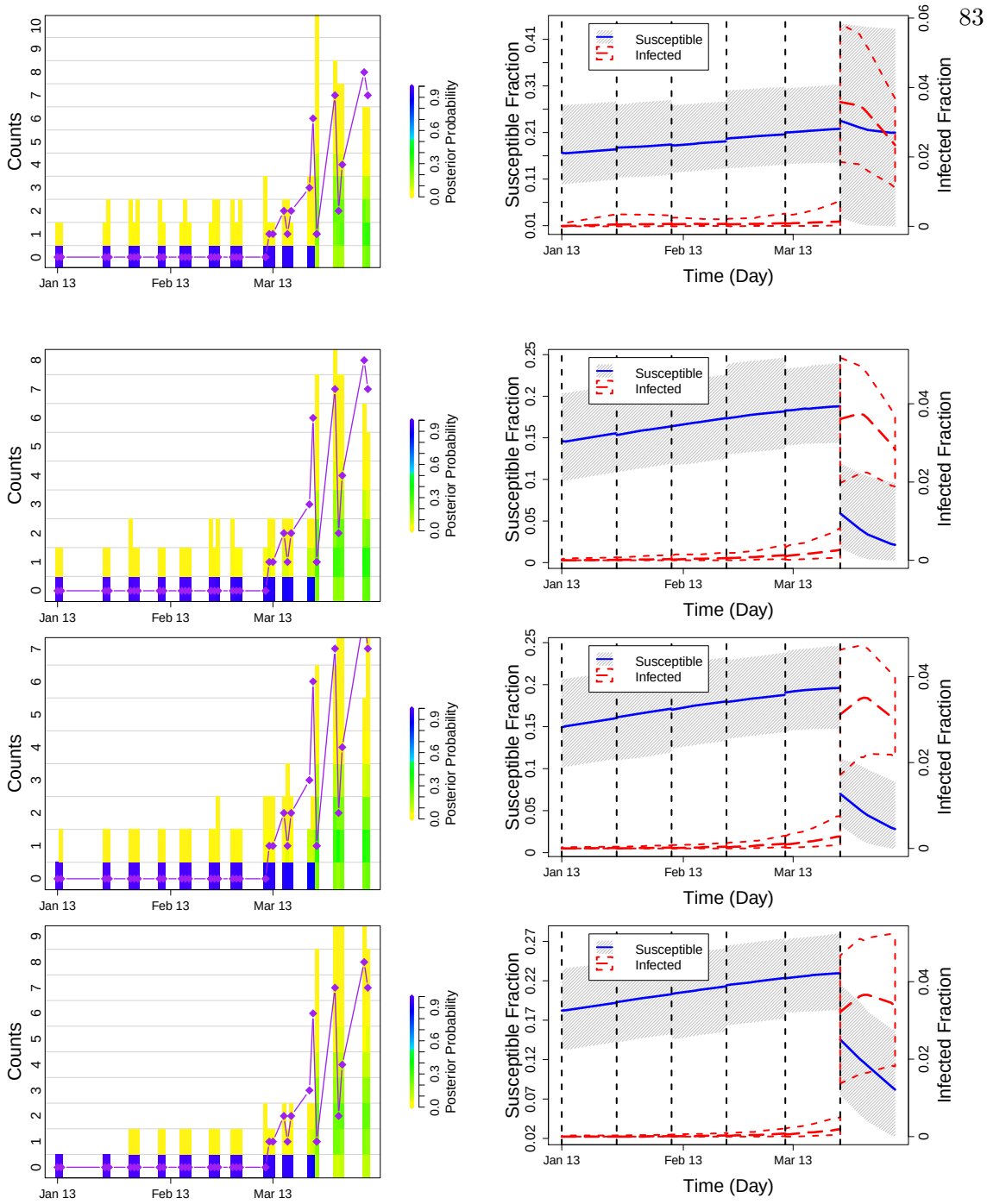

Figure 5.4: Summary of prediction results for the 2013 epidemic peak in Mathbaria; each row shows prediction results under different normal prior distributions on α_k parameters for $k = 0, \ldots, 15$. From top to bottom, values for the prior standard deviations are 5, 2, 1, and 0.5. Plots on the left compare the posterior probability of the predicted counts to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The plots on the right show how the trajectory of the predicted hidden fraction of susceptible and infected individuals change over the course of the epidemic. The gray area and the solid line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of infected individuals.

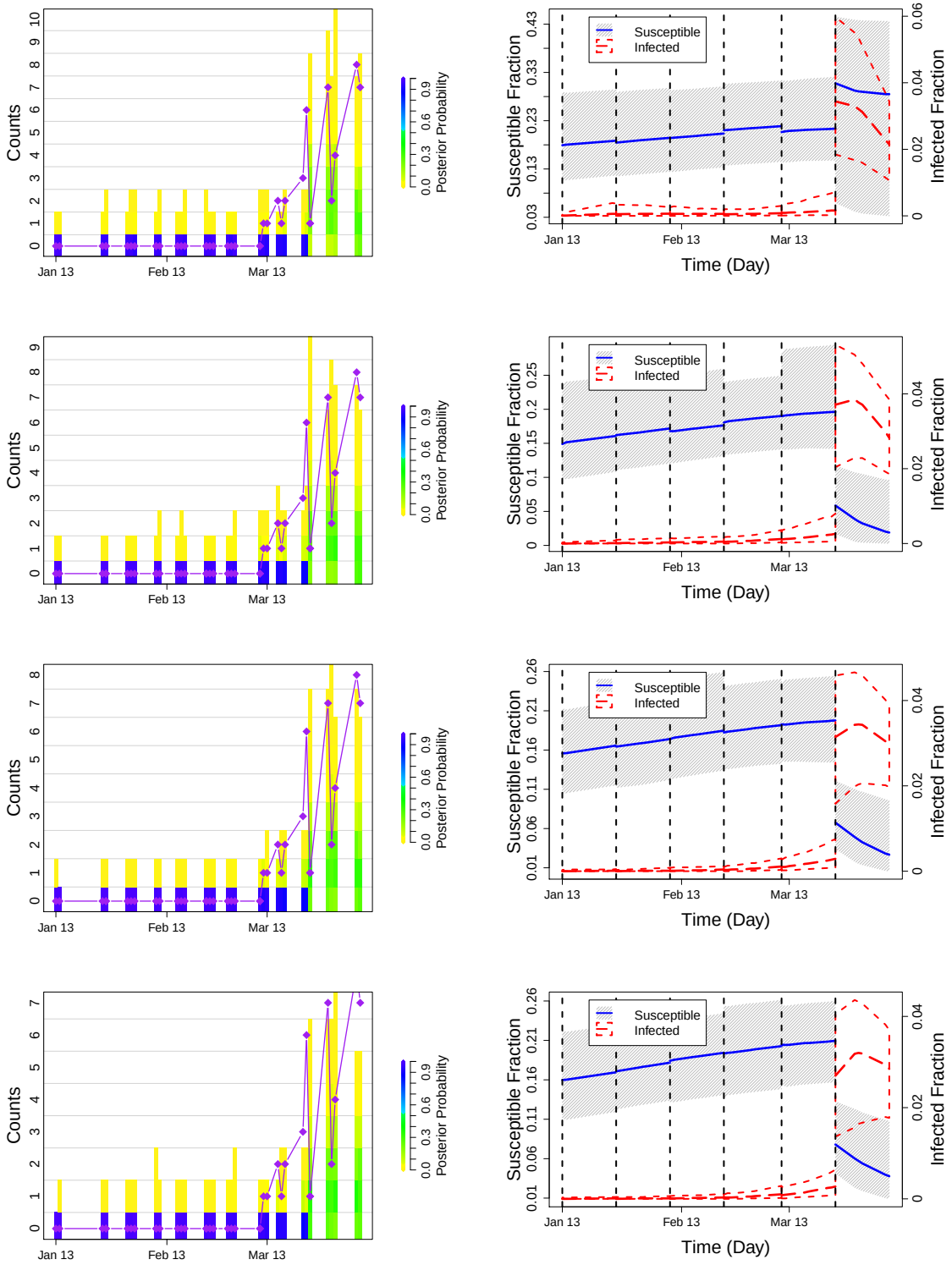

Figure 5.5: Summary of prediction results for the 2013 epidemic peak in Mathbaria; each row shows prediction results under different non-standardized t-distribution priors on α_k parameters for $k = 0, \ldots, 15$. From top to bottom, values for the prior standard deviations are 3.8, 1.55, 0.71, and 0.38. Plots on the left compare the posterior probability of the predicted counts to the test data (diamonds connected by straight line). The plots on the right show how the trajectory of the predicted hidden fraction of susceptible and infected individuals change over the course of the epidemic. See Figure 5.4 for details.

Chapter 6

MODELING ENVIRONMENTAL CONTRIBUTION TO CHOLERA OUTBREAKS USING A LATENT WATER COMPARTMENT

6.1 Hidden SIWR model

In the hidden SIRS model, we assume that the hazard rate of infection is $\beta I_t + \alpha(t)$ for each time t, where β is the infectious contact rate between infected individuals and susceptible individuals and $\alpha(t)$ is the time-varying environmental force of infection. In that model, infectious contact incorporates both direct person-to-person transmission of cholera and consumption of contaminated water. We now separate these contributions to transmission from infected individuals and explore models which incorporate a feedback loop from the infected individuals back into the environment to capture the effect of infected individuals excreting *V. cholerae* into the environment.

To accomplish this, we add a water compartment, W, that quantifies the concentration of *V. cholerae* in the environment. Instead of using an environmental force of infection, we incorporate the environmental covariates using the same function, $\alpha(t)$, as the rate of seasonal increase in water *V. cholerae* concentration. This SIWR model is similar to the SIWR model of Tien and Earn [2010] and Eisenberg et al. [2013b], but unique in the way it incorporates the environmental covariates. Transition rates for this SIWR compartmental model of disease transmission are shown in Figure 6.1. The hazard rate of infection is $\beta_I I_t + \beta_W W_t$ for each time t, where β_I represents the infectious contact rate between infected individuals and susceptible individuals and β_W represents force of infection from contact with or consumption of contaminated water. Infected individuals excrete *V. cholerae* into the environment/water compartment at rate κ. The time-varying function $\alpha(t)$ also contributes to the increase of the *V. cholerae* concentration in the water compartment. This concentration decays at rate η. Again, infected individuals recover from infection at rate γ, and recovered individuals lose immunity to infection and become susceptible at rate μ.

We model $\boldsymbol{X}_t = (S_t, I_t, R_t, W_t)$ as an inhomogeneous Markov process [Taylor and Karlin, 1998] with infinitesimal rates

$$\lambda_{(S,I,R,W),(S',I',R',W')}(t) = \begin{cases} (\beta_I I + \beta_W W)\, S & \text{if } S' = S-1,\ I' = I+1,\ R' = R,\ W' = W, \\ \gamma I & \text{if } S' = S,\ I' = I-1,\ R' = R+1,\ W' = W, \\ \mu R & \text{if } S' = S+1,\ I' = I,\ R' = R-1,\ W' = W, \\ \kappa I + \alpha(t) & \text{if } S' = S,\ I' = I,\ R' = R,\ W' = W+1, \\ \eta W & \text{if } S' = S,\ I' = I,\ R' = R,\ W' = W-1, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

where $\boldsymbol{X} = (S, I, R, W)$ is the current state and $\boldsymbol{X}' = (S', I', R', W')$ is a new state. We do not keep track of the number of recovered individuals because $R_t = N - S_t - I_t$.

The water compartment has no scale; it is used to quantify water contamination but may not necessarily be the exact amount of *V. cholerae* in the water. In fact, for a constant c the dynamics of the process are invariant if one makes the change of variables $c\beta_W$ and $(\kappa I + \alpha(t))/c$. This also controls the range of W, which can speed up simulation.

Again, we assume that $\boldsymbol{X}_t = (S_t, I_t, R_t, W_t)$ is not directly observable. Instead, we only observe y_t, the number of observed infections at time t, and assume y_t has a binomial distribution with size I_t, the number of infected individuals at time t, and success probability ρ, the probability of infected individuals seeking treatment; thus,

$$\Pr(y_{t_i} | I_{t_i}, \rho) = \binom{I_{t_i}}{y_{t_i}} \rho^{y_{t_i}} (1-\rho)^{I_{t_i} - y_{t_i}}. \tag{6.2}$$

We now have a hidden SIWR model similar to our hidden SIRS model, and we use the same Bayesian framework to estimate the parameters β_I, γ, κ, η, ρ, and the $k+1$ parameters associated with $\alpha(t)$, the time-varying seasonal growth rate. We assume $\alpha(t) = \exp(\alpha_0 + \alpha_1 C_1(t) + \cdots + \alpha_k C_k(t))$, where $C_1(t), \ldots, C_k(t)$ denote the k time-varying environmental covariates. Independent Poisson initial distributions are used for S_{t_0}, I_{t_0}, and

W_{t_0}, with means ϕ_S, ϕ_I, and ϕ_W respectively. Thus

$$\begin{aligned}\Pr(\boldsymbol{X}_{t_0}|\phi_S,\phi_I,\phi_W) &= \Pr(S_{t_0}|\phi_S)\times\Pr(I_{t_0}|\phi_I)\times\Pr(W_{t_0}|\phi_W)\\ &= \frac{\phi_S^{S_{t_0}}\exp(-\phi_S)}{S_{t_0}!}\times\frac{\phi_I^{I_{t_0}}\exp(-\phi_I)}{I_{t_0}!}\times\frac{\phi_W^{W_{t_0}}\exp(-\phi_W)}{W_{t_0}!}.\end{aligned}$$

In this Bayesian analysis, parameters β_W, ϕ_S, ϕ_I, ϕ_W, and μ are assumed to be known. Log transformations are used for parameters that are constrained to be greater than zero, such as β_I and γ and a logit transformation is used for the probability ρ. Transformed parameter values have independent Normal prior distributions that incorporate biological information if possible.

We are interested in the posterior distribution $\Pr(\boldsymbol{\theta}|\boldsymbol{y}) \propto \Pr(\boldsymbol{y}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta})$, where $\boldsymbol{y} = (y_{t_0},\ldots,y_{t_n})$, $\boldsymbol{\theta} = (\log(\beta_I),\log(\gamma),\mathrm{logit}(\rho),\alpha_0,\ldots,\alpha_k)$, and

$$\Pr(\mathbf{y}|\boldsymbol{\theta}) \;=\; \sum_{\boldsymbol{x}}\left(\prod_{i=0}^{n}\Pr(y_{t_i}|I_{t_i},\rho)\left[\Pr(\boldsymbol{X}_{t_0}|\phi_S,\phi_I,\phi_W)\prod_{i=1}^{n}p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_{i-1}},\boldsymbol{\theta})\right]\right).$$

Here $p(\boldsymbol{X}_{t_i}|\boldsymbol{X}_{t_{i-1}},\boldsymbol{\theta})$ for $i = 1,\ldots,n$ are the transition probabilities of the SIWR continuous-time Markov chain (CTMC). Again, this likelihood is intractable; there is no practical method to compute the finite time transition probabilities of the SIWR CTMC because the size of state space of (S_t, I_t) grows on the order of N^2 and the size of the the state space of W_t is infinite. Thus, we again use a particle marginal Metropolis-Hastings (PMMH) algorithm [Andrieu et al., 2010] to generate samples from the posterior distribution, which implements a sequential Monte Carlo algorithm to generate an unbiased estimate of the likelihood [Doucet et al., 2001].

In dealing with this more complex hidden Markov model, it is important to consider that certain parameters may not be identifiable. Analyzing the SIWR model of Tien and Earn [2010], Eisenberg et al. [2013b] found that, with incidence data, certain parameter combinations are structurally identifiable, such as $N\beta_I$, $N\beta_W\kappa$, and $N\rho$. However, whether these parameter combinations are practically identifiable remains an open question. Moreover, the SIWR model studied by Eisenberg et al. [2013b] did not include the seasonally varying growth rate $\alpha(t)$, which may create additional identifiability problems. Therefore,

we conduct a simulation study to examine what parameter combinations may be identifiable in our hidden SIWR model.

6.2 Simulating inhomogeneous SIWR

We use a tau-leaping algorithm [Gillespie, 2001, Cao et al., 2005] to simulate our SIWR CTMC by jumping over a small amount of time τ and approximating the number of events that happen in this time using a series of Poisson distributions. We define the rate functions $h_1(\boldsymbol{X}_t) = (\beta_I I_t + \beta_W W_t)$, $h_2(\boldsymbol{X}_t) = \gamma I_t$, and $h_3(\boldsymbol{X}_t) = \mu R_t$, $h_4(\boldsymbol{X}_t) = \kappa I_t + \alpha(t)$, $h_5(\boldsymbol{X}_t) = \eta W_t$, corresponding to the infinitesimal rates of the SIWR CTMC. Then $k_1 \sim \text{Poisson}(h_1(\boldsymbol{X}_t)\tau)$ represents the number of infections in time $[t, t+\tau)$, $k_2 \sim \text{Poisson}(h_2(\boldsymbol{X}_t)\tau)$ represents the number of recoveries in time $[t, t+\tau)$, $k_3 \sim \text{Poisson}(h_3(\boldsymbol{X}_t)\tau)$ represents the number of people that become susceptible to infection in time $[t, t+\tau)$, $k_4 \sim \text{Poisson}(h_4(\boldsymbol{X}_t)\tau)$ represents the increase in water *V. cholerae* concentration in time $[t, t+\tau)$, and $k_5 \sim \text{Poisson}(h_5(\boldsymbol{X}_t)\tau)$ represents the decay in water *V. cholerae* concentration in time $[t, t+\tau)$. We assume that the time-varying seasonal growth rate, $\alpha(t)$, remains constant each day and that $\tau = 1$ day, so our rates now remain constant within each tau jump. We use the modified tau-leaping algorithm, discussed in Section 2.2.4, to avoid simulating negative population sizes in our compartments. If the water compartment gets too large, simulation of the SIWR CTMC involves a prohibitively long computing time, since the modified tau-leaping algorithm uses a single step algorithm (like the Gillespie algorithm) when one of the compartments has a low population size. To control the size of W in the SMC algorithm, we set β_W in the PMMH algorithm.

6.3 Bayesian analysis of data simulated from SIWR model

We simulate from the hidden SIWR model using a population size of $N = 5000$ and assume independent Poisson initial distributions for S_{t_0}, I_{t_0}, and W_{t_0}, with means $\phi_S = 1400$, $\phi_I = 16$, and $\phi_W = 50$. The other parameters are set at $\beta_I = 2.144 \times 10^{-5}$, $\beta_W = 7 \times 10^{-6}$, $\alpha_0 \approx 0.39$, $\alpha_1 = 3.5$, $\gamma = 0.1$, $\kappa = 0.02$, $\eta = 1/30 \approx 0.03$, and $\mu = 0.0009$. Rates are measured in the number of events per day. Using the modified Gillespie algorithm to simulate from the SIWR model, as described in 2.2.2, the resulting (S_t, I_t, W_t) chain is given

in Figure 6.2. From the hidden data, we simulate the observed number of infections, y_t, as $y_t \sim \text{Binomial}(I_t, \rho)$, where $\rho = 0.05$. Here case observations occur once every two weeks.

We use diffuse Normal prior distributions on α_0 and α_1, both centered at -6 and with standard deviations of 5. The transformed parameter value $\log(\beta_I)$ has a Normal prior centered at $\log(1 \times 10^{-4})$ and with a standard deviation of 3. Normal prior distributions for $\log(\kappa)$ and $\log(\eta)$ are centered at $\log(0.005)$ and $\log(0.01)$, both with standard deviations of 1.5. For $\text{logit}(\rho)$, the Normal prior distribution is centered at $\text{logit}(0.01)$ and has a standard deviation of 2. The Normal prior for the transformed parameter $\log(\gamma)$ is centered at $\log(0.1)$ with a standard deviation of 0.09.

For estimation, the population size is set at 5000 and the loss of immunity rate μ is set at 0.0009. We compare results from models with different assumptions about the means of the Poisson initial distributions for S_{t_0}, I_{t_0}, and W_{t_0}. Values for ϕ_S/N, ϕ_I/N, and ϕ_W are set above the true values, (0.38, 0.0064, 70), at the truth (0.28, 0.0032, 50), below the true values (0.18,0.0016, 30), and further below (450, 4, 10). In all of these runs, β_W is set at its true value. The PMMH algorithm starts with a burn-in run of 10000 iterations, a secondary run of 10000 iterations, and a final run of 50000 iterations. We save only every 10th iteration to thin the chains, and we use $K = 100$ particles in the SMC algorithm. Marginal posterior distributions for the parameters of the SIWR model from the final runs of the PMMH algorithms are shown in Figure 6.3. Overall, posterior distributions are very similar regardless of assumptions about ϕ_S/N, ϕ_I/N, and ϕ_W, especially for parameters γ, α_0, α_1, and ρ. When the model parameters are misspecified, the ranges of the posteriors for β_I and κ are large, and the true value of η is in the tail of its posterior distribution. Trace plots and autocorrelation plots for the parameters, from the model that assumes ϕ_S/N, ϕ_I/N, and ϕ_W are set at the truth, are shown in Figure 6.4.

6.3.1 Prediction results

We test the predictive ability of this model using the techniques described in Chapter 3. We break our data into multiple training and test sets, using the cut-off days shown in Figure 6.2. For each of the four cut-off days, a training set of data, which includes data from the

first observation up until the cut-off day, is used to approximate the posterior distribution of the parameters of the SIWR model. Then parameters are sampled from this distribution to simulate possible future states of the hidden SIWR system and future observed case counts until the next cut off 28 days later. We compare these to the test data to see how well the SIWR model predicts cholera outbreaks.

Figure 6.5 shows prediction summaries from models with values for ϕ_S/N, ϕ_I/N, and ϕ_W set over the true values, (0.38, 0.0064, 70), at the truth (0.28, 0.0032, 50), under the true values (0.18,0.0016, 30), and further under (450, 4, 10). Predictions are similar for all sets of parameter assumptions; this is not surprising, as the posterior distributions of the parameters are also similar. The posterior distributions of predicted counts for all settings encompass the test data well. The ranges of the 95% quantiles of the predictive distributions for the fractions of susceptible and infected individuals are largest when ϕ_S/N, ϕ_I/N, and ϕ_W are set above the true values, and they decrease as the assumed values for these parameters decrease.

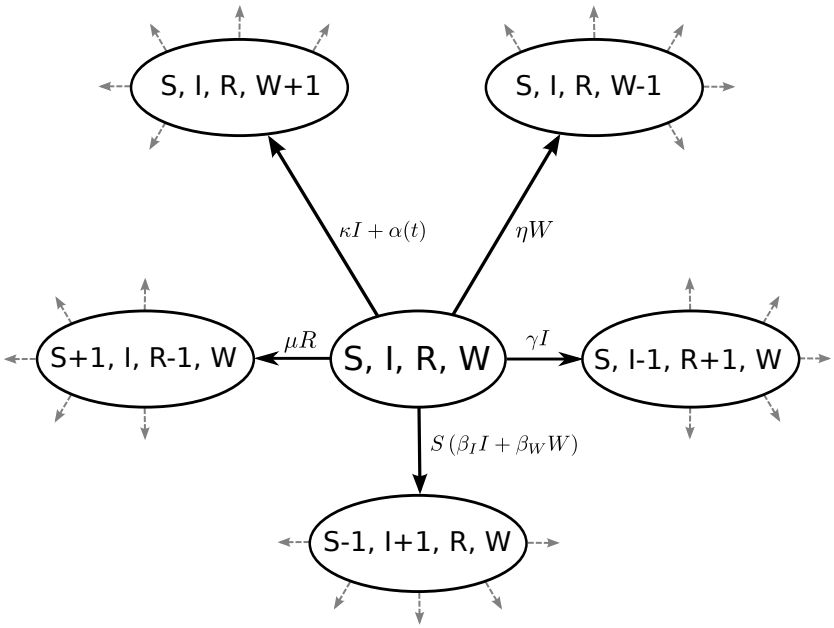

Figure 6.1: State transitions for the SIWR model for cholera. S, I, and R denote the numbers of susceptible, infected, and recovered individuals, and W quantifies the concentration of *V. cholerae* in the environment. The state $\boldsymbol{X} = (S, I, R, W)$ can transition to one of five new states corresponding to a susceptible becoming infected, an infected recovering from infection, a recovered individual losing immunity to infection and becoming susceptible, the water compartment increasing, or the water compartment decreasing. The parameter β_I is the infectious contact rate, β_W is the rate of infection from contact with contaminated water, κ is the rate at which infected individuals excrete *V. cholerae* into the environment, $\alpha(t)$ is the seasonal increase in water *V. cholerae* concentration, η is the rate of decay in water *V. cholerae* concentration, γ is the recovery rate, and μ is the rate at which immunity is lost.

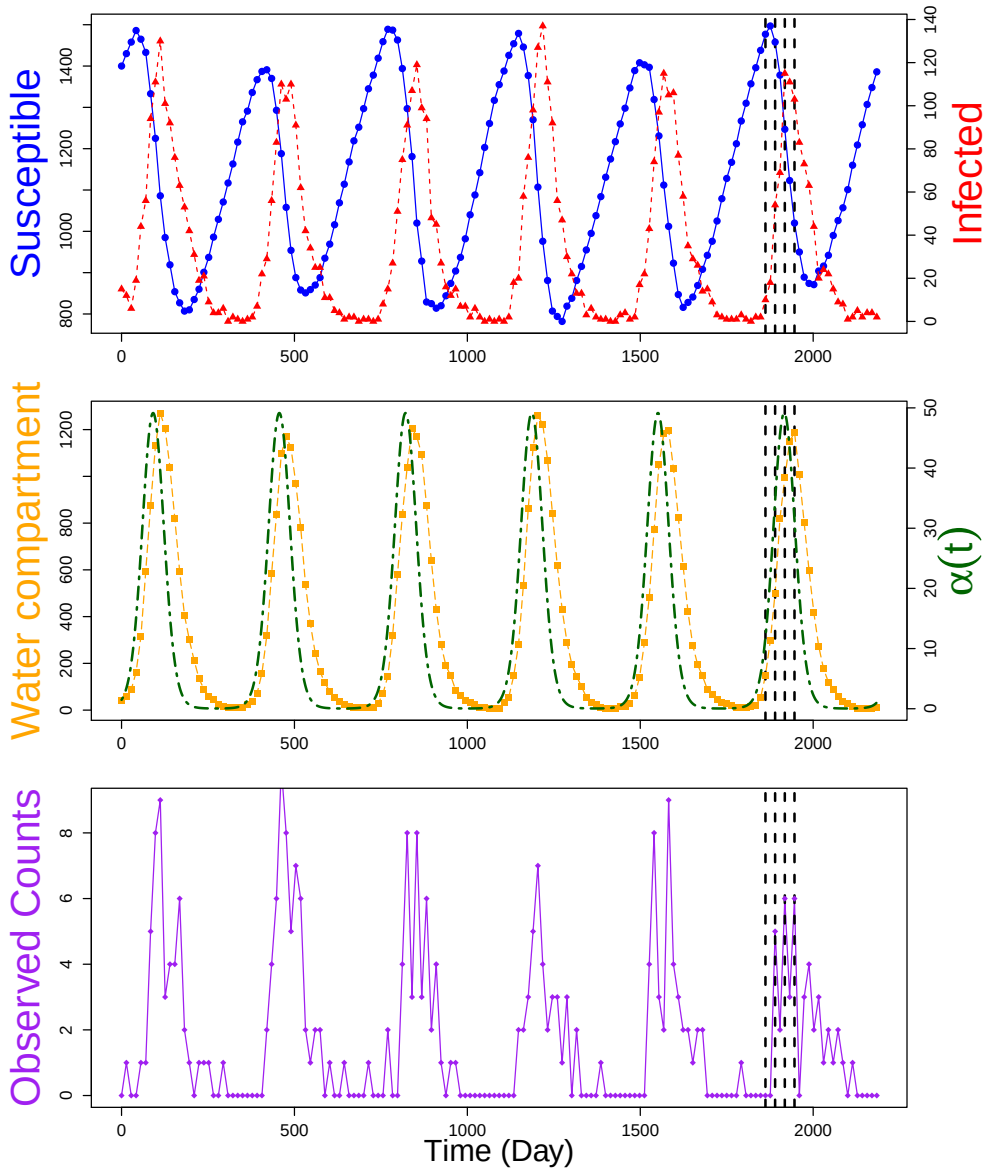

Figure 6.2: Plots of simulated hidden states (counts of susceptible individuals, S_t, infected individuals, I_t, and the water compartment, W_t) and the observed data ($\alpha(t)$ and $y_t \sim \text{Binomial}(I_t, \rho)$ = number of observed infections) plotted over time (t). The top plot shows the dynamics of the number of susceptible and infected individuals over time. The middle plot shows the dynamics of the water compartment and rate of seasonal increase in water *V. cholerae* concentration, $\alpha(t)$. There is a slight delay in the increase in the water compartment after $\alpha(t)$ increases, and the decay of the water compartment is more gradual than the decrease in $\alpha(t)$. The dashed vertical black lines represent cut offs between the training sets and test data.

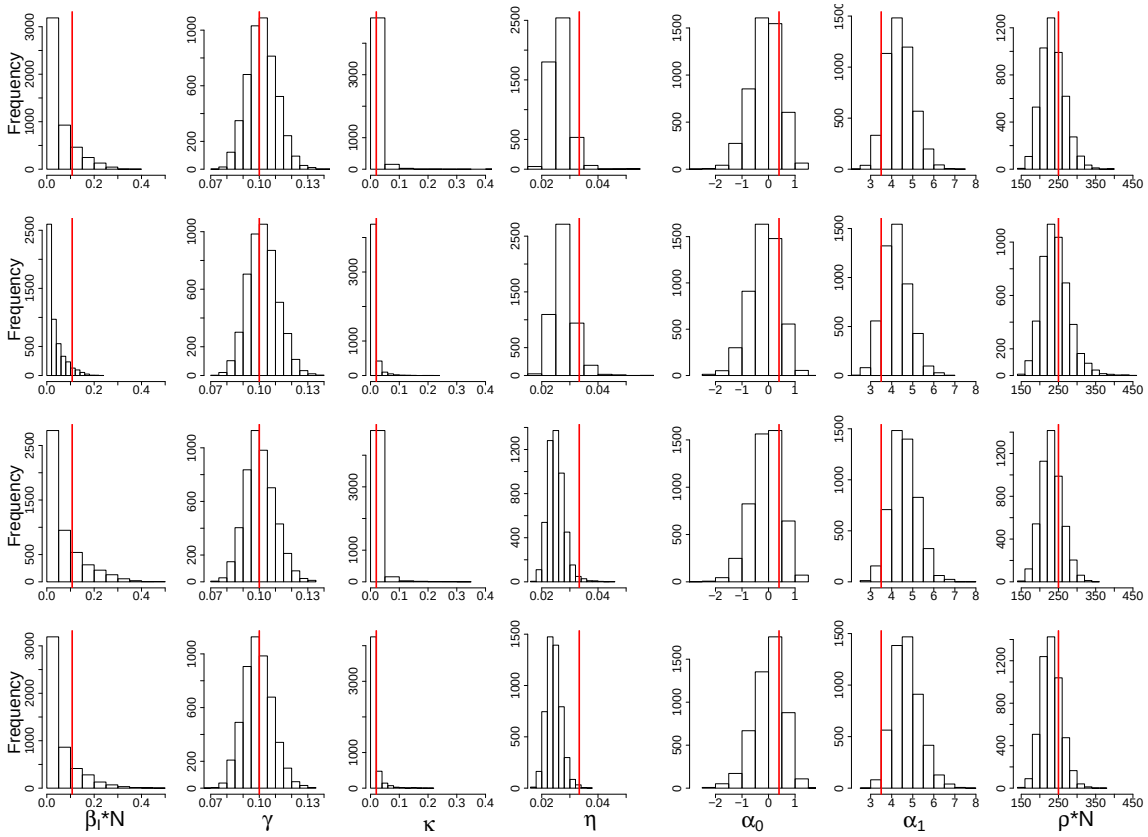

Figure 6.3: Marginal posterior distributions for the parameters of the SIWR model from the final run of the PMMH algorithm. The red lines denote the true values of the parameters. From top to bottom, assumed ϕ_S/N, ϕ_I/N and ϕ_W are set above the true values, (0.38, 0.0064, 70), at the truth (0.28, 0.0032, 50), below the true values (0.18,0.0016, 30), and further below (450, 4, 10). All marginal posterior distributions contain the true values in their ranges, and look similar under different parameter value assumptions.

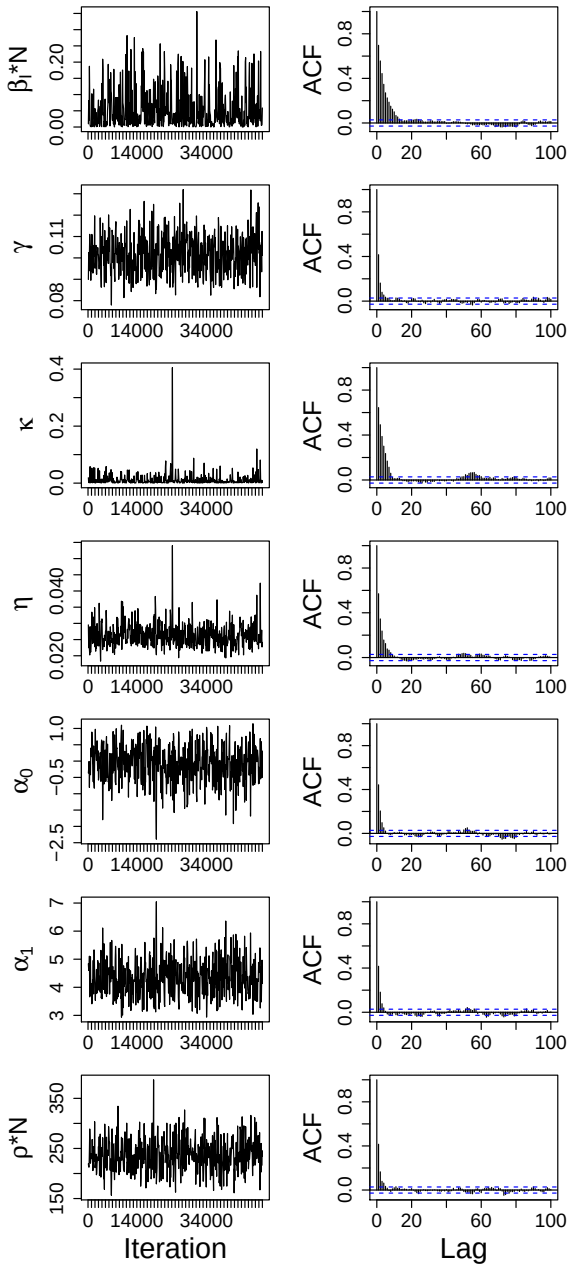

Figure 6.4: Trace plots and autocorrelation plots for the parameters of the SIWR model from the final run (50000 iterations) of the PMMH algorithm for simulated data. Trace plots are thinned to display only 500 iterations; autocorrelation plots are thinned to display only 5000 iterations. Parameters ϕ_S/N, ϕ_I/N and ϕ_W are set at the truth (0.28, 0.0032, 50).

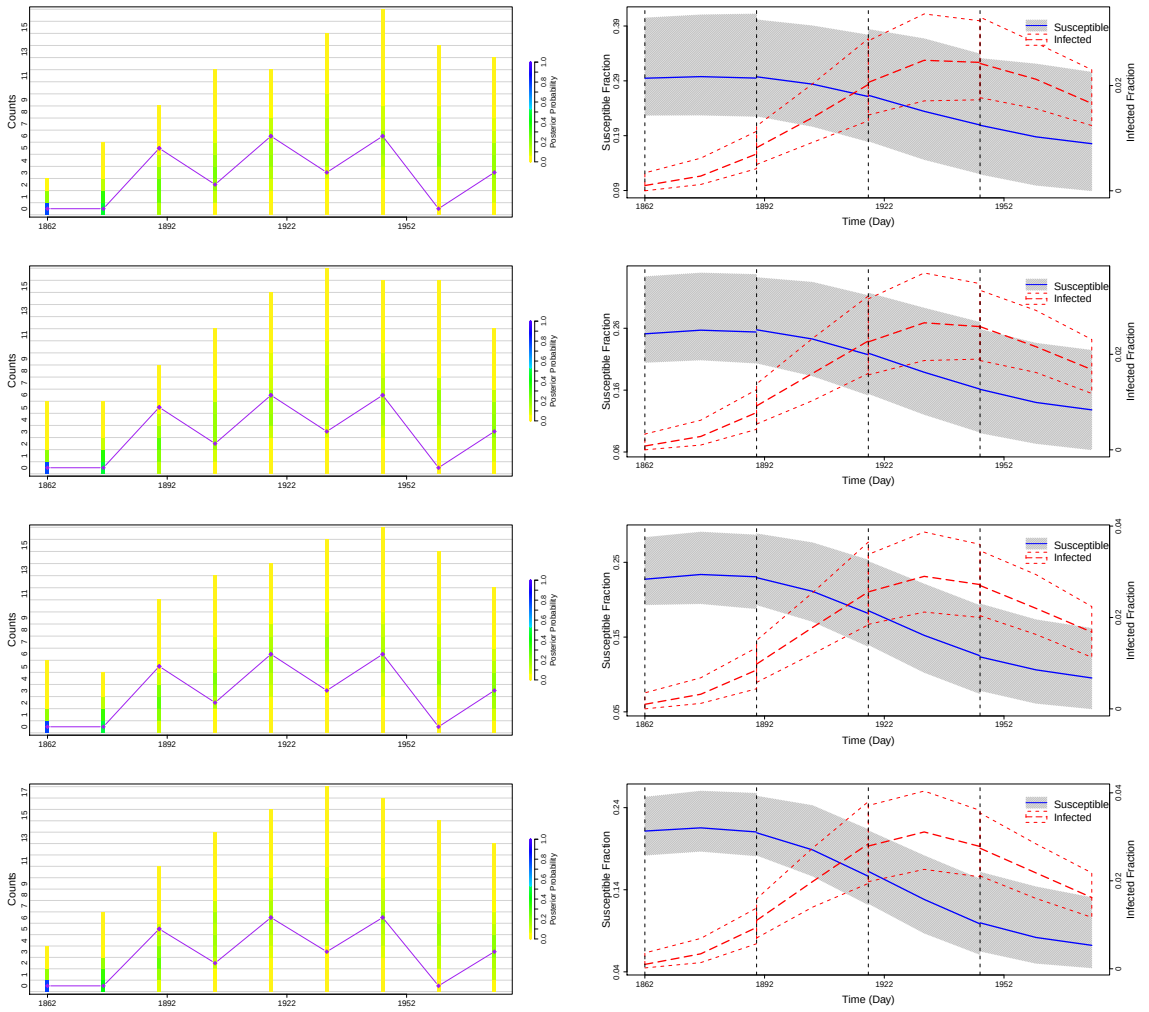

Figure 6.5: Summary of prediction results for simulated data; each row shows prediction results under different assumptions about the values of ϕ_s/N, ϕ_I/N, and ϕ_W. From top to bottom, values for ϕ_S/N, ϕ_I/N, and ϕ_W are set above the true values, (0.38, 0.0064, 70), at the truth (0.28, 0.0032, 50), below the true values (0.18,0.0016, 30), and further below (450, 4, 10). Plots on the left compare the posterior probability of the predicted counts to the test data (diamonds connected by straight line). The coloring of the bars is determined by the frequency of each set of counts in the predicted data for each time point. The plots on the right show how the trajectory of the predicted hidden fraction of susceptible and infected individuals change over the course of the epidemic. The gray area and the solid line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of susceptibles. The short dashed lines and the long dashed line denote the 95% quantiles and median, respectively, of the predictive distribution for the fraction of infected individuals. The timing of the epidemic is the same for all settings.

6.4 Using cholera incidence data and covariates from Mathbaria, Bangladesh

Using the water depth and water temperature covariates and cholera incidence data from Mathbaria, discussed in Section 3.6 and shown in Figure 6.6, we estimate the parameters of the hidden SIWR model. We again consider the daily time intervals $A_i := [i, i+1)$ for $i \in \{t_0, t_0+1, \ldots, t_n - 1\}$ and define the environmental force of infection $\alpha(t) = \alpha_{A_i}$ for $t \in A_i$ and $t \geq 21$ days where $\alpha_{A_i} = \exp\left[\alpha_0 + \alpha_1 C_{WD}(i-21) + \alpha_2 C_{WT}(i-21)\right]$. Here 21 days is the length of the covariate lag, and the covariates are the smoothed standardized daily values $C_{WT}(i) = (WT(i) - \overline{WT})/s_{WT}$ and $C_{WD}(i) = (WD(i) - \overline{WD})/s_{WD}$, where $\overline{X}$ is the mean of the measurements for all i and s_X is the sample standard deviation.

We again set the loss of immunity rate $\mu = 0.0009$, setting the average length of immunity, $1/\mu$, to 3 years [Sack et al., 2004]. Also, the population size N, which quantifies the size catchment area for the medical center, is assumed to be 10000. The means of the Poisson initial distributions for S_{t_0}, I_{t_0}, and W_{t_0} are set at $\phi_S/N = 0.08$, $\phi_I/N = 0.001$, and $\phi_W = 37.8$, based on posterior samples from a tuning run of the PMMH algorithm.

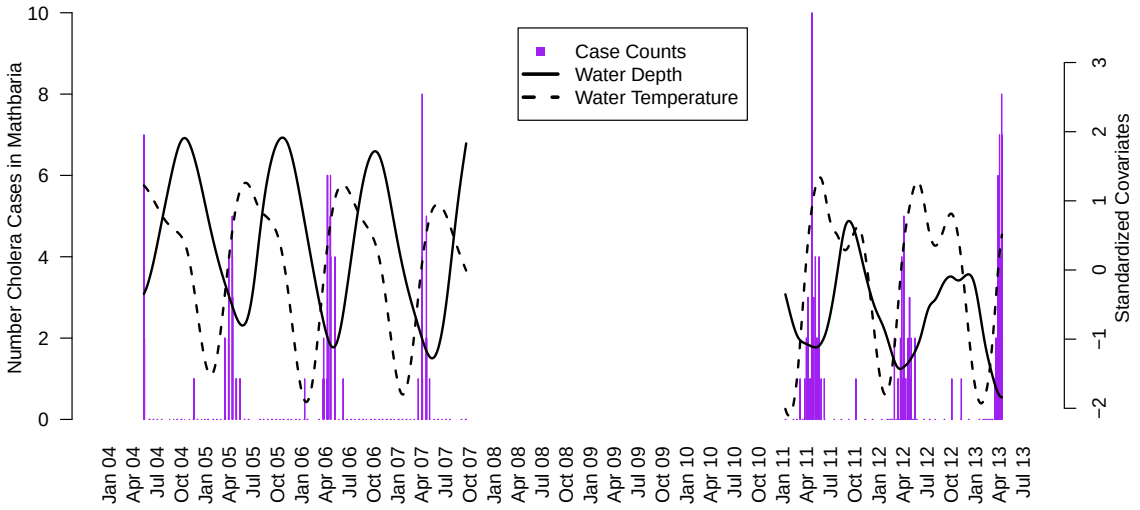

Figure 6.6: Barplot of cholera case counts in Mathbaria, Bangladesh and the standardized covariate measurements over time. The covariates are shown with a lag of three weeks. No data were collected from October 2007 through November 2010.

We use uninformative, diffuse Normal prior distributions on the time-varying environmental covariates α_1 and α_2, centered at 0 and with standard deviations of 5. The transformed parameter values $\log(\beta_I)$ and α_0 are also given diffuse Normal prior distributions, centered at $\log(1.25 \times 10^{-7})$ and -8, respectively, with standard deviations of 5. We incorporate biological information into the Normal prior distributions for $\log(\gamma)$ and $\text{logit}(\rho)$, assuming the prior for $\log(\gamma)$ is centered at $\log(0.1)$ with a standard deviation of 0.09 and the prior for $\text{logit}(\rho)$ is centered at $\text{logit}(0.0008)$ and has a standard deviation of 2.

The PMMH algorithm starts with a burn-in run of 400 iterations, a secondary run of 1600 iterations, and a final run of 32000 iterations. To thin the chains, we save only every fourth iteration. We use $K = 100$ particles in the SMC algorithm. Posterior medians and 95% credible intervals for the parameters of the SIWR model are given in Table 6.1. Figure 6.7 shows trace plots, autocorrelation plots, and posterior histograms, and Figure 6.8 shows bivariate scatterplots of parameters.

Similar to the results of Section 3.6, both water depth and water temperature have a significant relationship with water *V. cholerae* concentration since the credible intervals for α_1 and α_2 do not include zero. They also maintained the magnitude and direction of their relationship; decreasing water depth increases water *V. cholerae* concentration, and increasing water temperature increases water *V. cholerae* concentration. The credible interval for ρ in the SIWR model overlaps the credible interval for ρ in the SIRS model. The posterior median for η is much higher than previously used; Hartley et al. [2005] assumed this parameter should be 1/30.

We test the predictive ability of our model by dividing the data into staggered training and test sets using seven cut off days during the last epidemic peak. For each cut off time, we run the PMMH algorithm on a training set of data which includes all observed data up until the cut off time. Table 6.2 shows effective sample sizes for the parameters of the SIWR model from each of the seven PMMH algorithm outputs. For the fourth, fifth, and sixth cut off times, estimation using the training set of data took longer than 294 hours, so the final run of 32000 iterations was terminated early. The total number of iterations for each final run are reported in Table 6.2. The chains are not mixing very well, as seen in the small effective sample sizes. Timing is a problem in the current framework; the length of

Table 6.1: Posterior medians and 95% equitailed credible intervals for the parameters of the SIWR model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh.

Coefficient	Estimate	95% CIs
$\beta_I \times N$	0.88	(0.65 , 1.12)
γ	0.12	(0.10 , 0.14)
κ	0.0041	(0.0002, 0.0701)
η	1.01	(0.25 , 5.31)
α_0	3.15	(1.27 , 5.13)
α_1	-2.05	(-2.81 , -1.52)
α_2	2.34	(1.86 , 2.99)
$\rho \times N$	58.34	(47.02 , 72.48)

Table 6.2: Effective sample sizes for the parameters of the SIWR model estimated using clinical and environmental data sampled from Mathbaria, Bangladesh. Mixing is a problem, especially for the parameters κ, η, and α_0.

		Effective sample sizes							
Cut off	Iterations	β_I	γ	κ	η	α_0	α_1	α_2	ρ
1	32000	169	188	34	25	32	51	99	283
2	32000	241	329	59	34	100	108	281	387
3	32000	499	500	80	189	152	185	404	714
4	20000	270	366	60	61	64	287	319	532
5	24000	474	454	72	201	179	410	290	565
6	16000	37	264	11	12	14	111	191	290
7	32000	382	596	103	52	47	303	725	514

computing time is very unpredictable as simulation in the SMC algorithm can be very slow if W is large. These PMMH runs took from 80 to 294 hours.

6.4.1 Prediction results

Using the data from Mathbaria and cut-offs shown in Figure 6.9, we predict cholera dynamics during the 2013 cholera outbreak. This is a subset of the data predicted in Section 3.6. Prediction results look similar to those obtained using the hidden SIRS model. The quantiles of the hidden fractions of susceptible and infected individuals cover a similar range as the quantiles in Figure 3.12. For the SIWR model, the resetting in the quantiles between cut-

offs is likely due to the poor mixing. The posterior probability of the predicted counts in Figure 6.9 has slightly more mass on higher counts than we saw in Figure 3.12; for example, counts of five predicted cases occur more often. However, we are still not capturing the observed case counts in the test data.

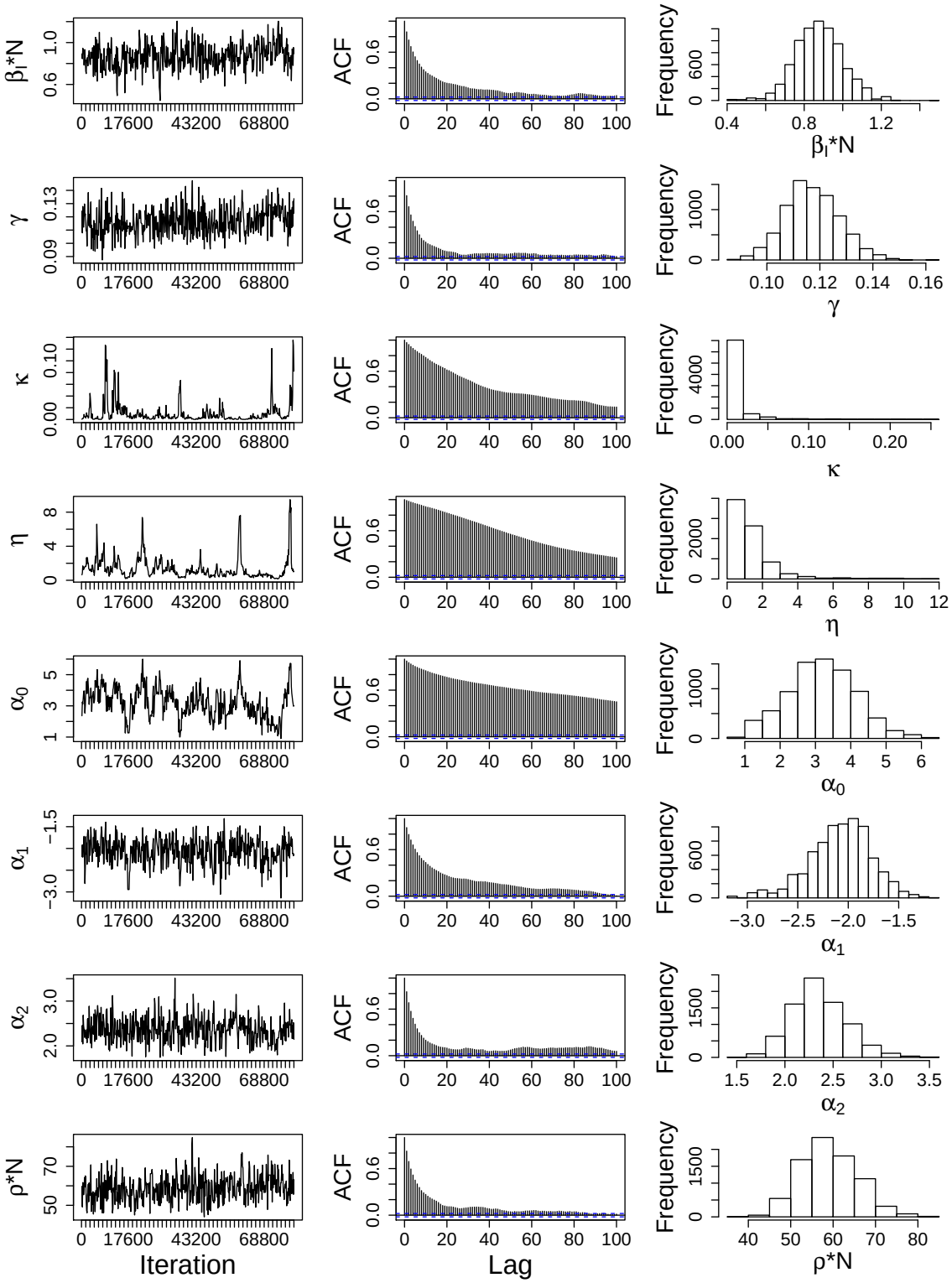

Figure 6.7: Summary plots of the PMMH algorithm output (final run of 32000 iterations) for the parameters of the SIWR model with data from Mathbaria, Bangladesh. Trace plots are thinned to display only 500 iterations; autocorrelation plots and posterior histograms are thinned to display 8000 iterations.

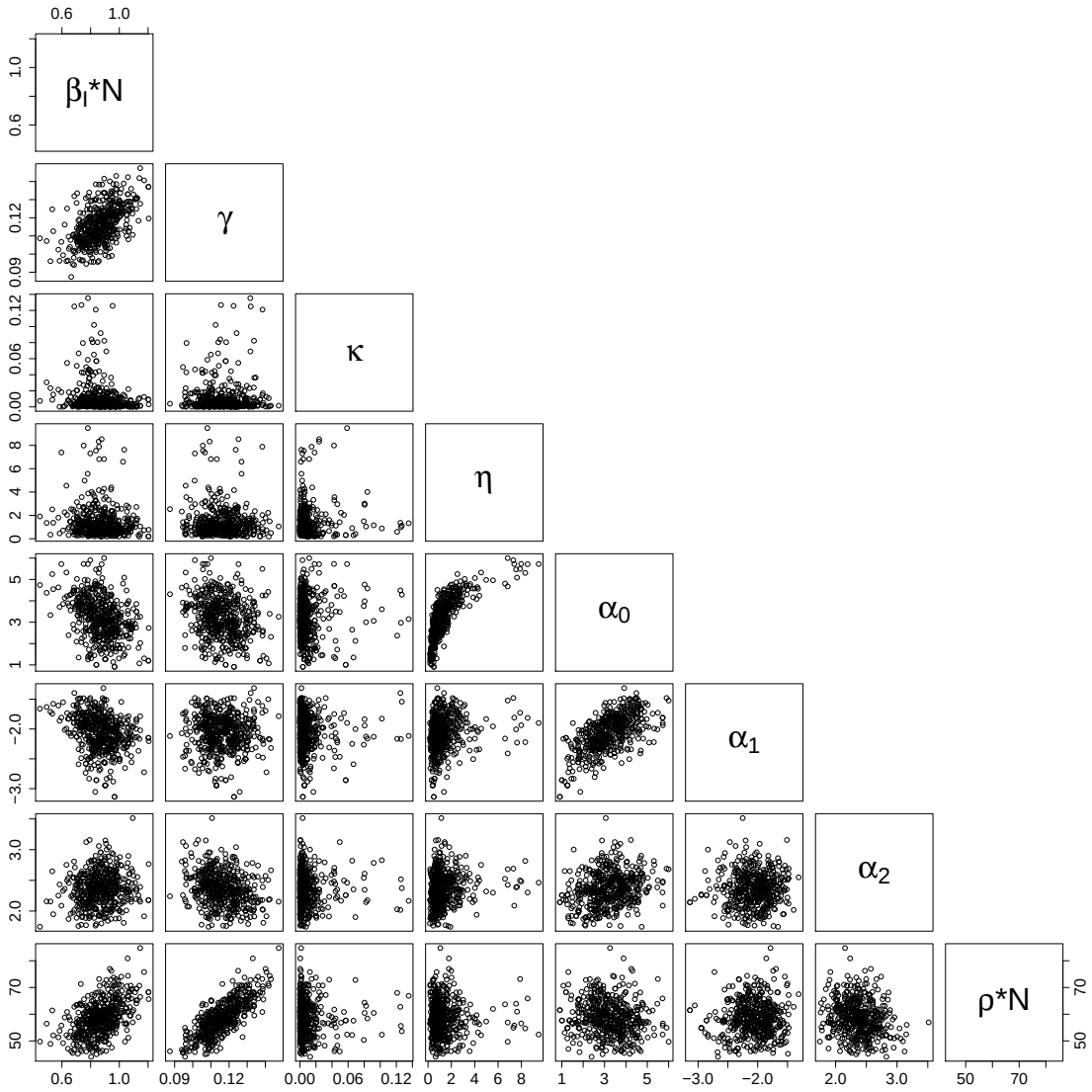

Figure 6.8: Bivariate scatterplots of parameters of the SIWR model estimated using data from Mathbaria, Bangladesh. Scatterplots are thinned to display only 500 samples, so only every 16th sample from the posterior distribution is plotted.

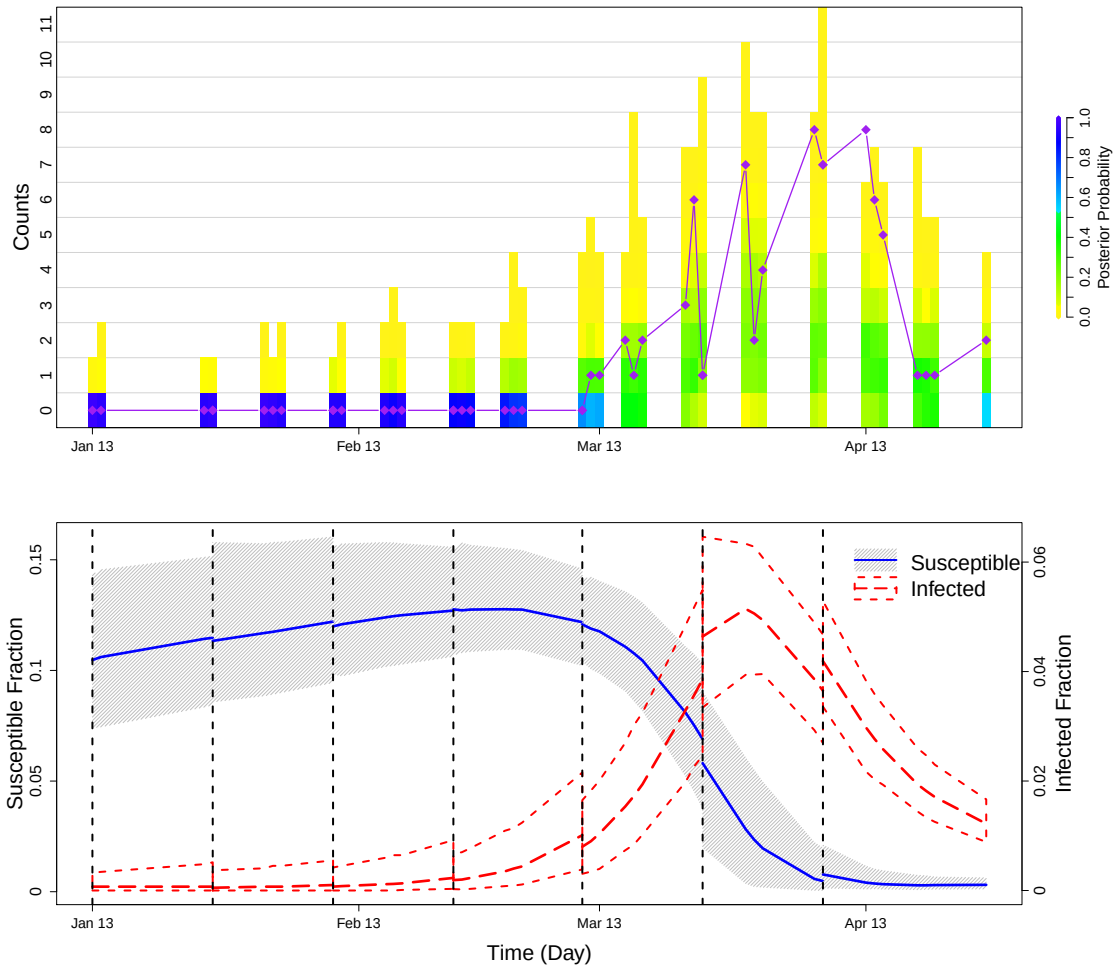

Figure 6.9: Summary of prediction results for the last epidemic peak in the Mathbaria data. We again run PMMH algorithms on training sets of the data, which are cut off at each of the dashed black lines in the bottom plot, and future cases are predicted until the next cut off. The top plot compares the posterior probability of the predicted counts to the test data (purple diamonds and line), and the bottom plot shows how the trajectory of the predicted hidden states changes over the course of the epidemic. See the caption of Figure 6.5 for more details.

6.5 Bayesian analysis using SIRS model on data simulated from an SIWR model

In the previous chapters, we assume the hidden SIRS model describes the dynamics of cholera outbreaks in Bangladesh. In this chapter, we assume a more biologically realistic model of cholera transmission, but encounter problems in estimating the parameters of the hidden SIWR model using data from Mathbaria. Using the hidden SIRS model is preferable as long as assuming this simplified version of cholera dynamics does not impact the validity of our predictions. We use simulated data to study the effects of misspecification of the data generating process on estimation and prediction.

Using the data simulated under the SIWR model, described in Section 6.3 and shown in Figure 6.2, we implement the PMMH algorithm under the assumption of a hidden SIRS model, as in Chapters 3, 4, and 5. We test the effects on parameter estimation and prediction when the model that we use to fit the data does not match the data generating process.

We assume the same settings for the PMMH algorithm that we used for the simulated data in the Chapter 3 SIRS analysis. We use uninformative, diffuse Normal prior distributions for $\log(\beta)$, α_0, and α_1, centered at $\log(0.000125)$, -8, and 0, respectively, and with standard deviations of 5. The Normal prior distribution for $\log(\gamma)$ is centered at $\log(0.1)$ with a standard deviation of 0.09, and the Normal prior for $\text{logit}(\rho)$ is centered at $\text{logit}(0.03)$ and has a standard deviation of 2. We assume the best case scenario for the parameters that are assumed known; the population size, ϕ_S and ϕ_I are set to their true values of 5000, 1400 and 16. Using this simulated data, the PMMH algorithm starts with a burn-in run of 30000 iterations, a secondary run of 20000 iterations, and a final run of 400000 iterations. To thin the chains, we save only every 10th iteration. We use $K = 100$ particles in the SMC algorithm.

Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using data simulated from an SIWR model are shown in Table 6.3, along with true values for parameters that are comparable between the two models. Figure 6.10 shows summary plots of the PMMH algorithm output. The true value of $\rho \times N$ from the SIWR model is in the credible interval for the corresponding parameter of the SIRS model.

Table 6.3: Posterior medians and 95% equitailed credible intervals for the parameters of the SIRS model estimated using data simulated from an SIWR model. True values for parameters that are comparable between the two models are shown.

SIRS Parameter	SIWR True value	Estimate	95% CIs
$\beta \times N$	—	0.2852	(0.2066, 0.3809)
γ	0.1	0.09	(0.07 , 0.1)
α_0	—	-10.56	(-12.27 , -9.39)
α_1	—	4.27	(2.8 , 6.2)
$\rho \times N$	250	229.51	(180.74, 292.2)

Surprisingly, the true value of $\alpha_1 = 3.5$ for the SIWR model is in the credible interval for the parameter α_1 from the SIRS model. This means that the covariate effect is still being captured in this simulation, regardless of how we model the relationship of environmental covariates to cholera outbreaks. The posterior estimate of α_0 seems to compensate for the model misspecification, as the true value of $\alpha_0 = 0.39$ for the SIWR simulated data is outside the credible interval for the SIRS α_0 parameter. The posterior distribution for γ also seems to be shifted away from the truth, though the true value is still inside the credible interval. The chains are mixing well, as seen in the trace plots in Figure 6.10. Effective sample sizes range from 14022 to 15815.

6.5.1 Prediction results

We test the predictive ability of the SIRS model on data generated from an SIWR model using staggered training and test sets of data, as described in Section 6.3.1. Despite the model misspecification, the model predictions look good, as seen in Figure 6.11. The trajectory of the predicted fraction of infected is similar in both Figure 6.5 and Figure 6.11, however the quantiles for the predicted fraction of susceptible individuals is consistently higher for predictions made from the misspecified model. The range of the predicted count posterior probabilities looks very similar to the ranges seen in Figure 6.5. For predictions from the misspecified model, higher predicted case counts appear earlier in the epidemic, as seen in the green coloring for counts of 2 and 3 at the first and second observation time.

The distribution of predicted counts also seems more spread over the counts at all observation times. However, the general trend and important features of the epidemic curve are captured.

6.6 Discussion

We have laid a promising foundation for the inclusion of a latent water compartment in the Bayesian analysis of cholera outbreaks. Using simulated data, we are able to recover the true parameters of the SIWR model and predict future simulated cholera outbreaks. For the analysis of the data from Mathbaria, we encountered issues within the SMC algorithm for some sets of proposed parameter values. Future work will look at ways to constrain the size of the water compartment. This is currently a major constraint on estimation, as large values of W slow down simulation in the SMC algorithm. Eisenberg et al. [2013b] found that, with data relating to the concentration of *V. cholerae* in the water, parameter identifiability in the SIWR model improved. Adding this type of data to our emission probability may help in our model.

We find that we can still predict outbreaks well when we fit the parameters of the SIRS model to data generated by the SIWR model. This suggests that the predictions made with the SIRS model using the data from Mathbaria in Section 3.6.2 may not be too far off, even if the model is not biologically realistic.

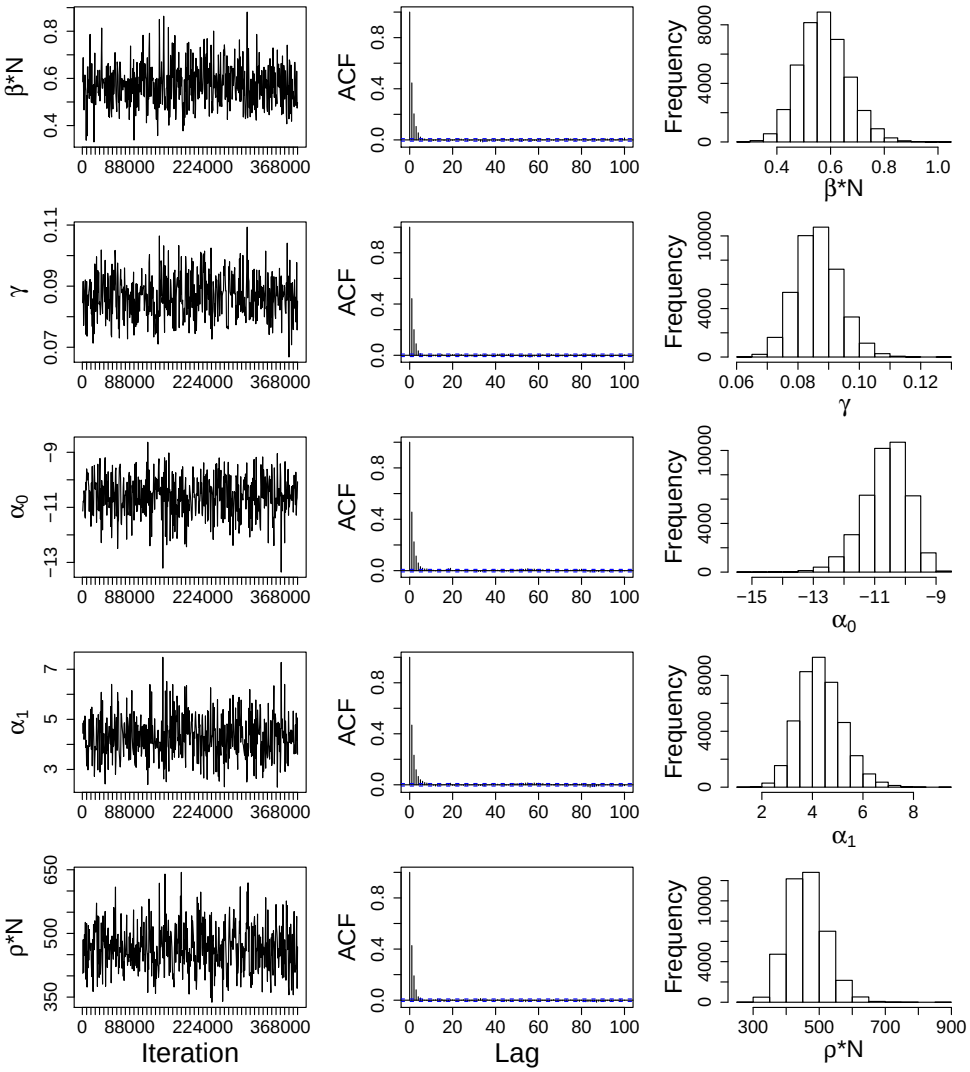

Figure 6.10: Summary plots of the PMMH algorithm output (final run of 400000 iterations) for the parameters of the SIRS model estimated using data simulated from an SIWR model. Trace plots are thinned to display only 500 iterations; autocorrelation plots and posterior histograms are thinned to display 40000 iterations.

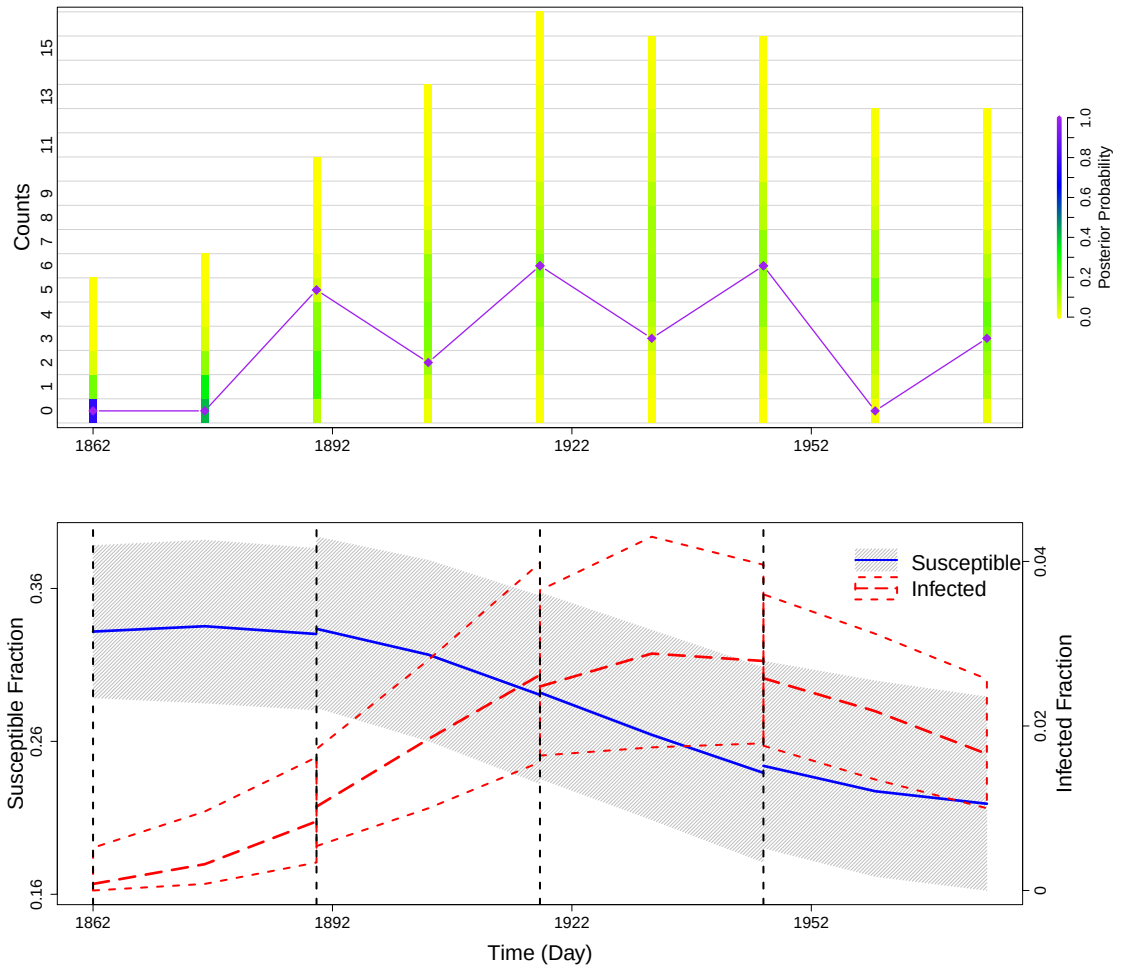

Figure 6.11: Summary of prediction results for data simulated from an SIWR model. We approximate the posterior distribution of the parameters of the hidden SIRS model using training sets of the data, which are cut off at each of the dashed black lines in the bottom plot, and future cases are predicted until the next cut off. The top plot compares the posterior probability of the predicted counts to the test data (purple diamonds and line), and the bottom plot shows how the trajectory of the predicted hidden states changes over the course of the epidemic. See the caption of Figure 6.5 for more details.

Chapter 7

FUTURE DIRECTIONS

We have used a Bayesian framework to estimate the parameters in two nonlinear dynamical models for cholera transmission in Bangladesh. We used a particle marginal Metropolis-Hastings algorithm to sample from the posterior distribution of the unobserved data and parameters given the observed cholera case counts. Using SMC to generate an estimate of the likelihood allowed us to use a Metropolis-Hastings approach without needing to explicitly calculate the problematic transition probabilities of our SIRS and SIWR models. We tested this method using both simulated data and data from Mathbaria, Bangladesh, then extended this framework to analyze data sampled from multiple areas in Bangladesh and to include many covariates. The following sections describe plans for further extensions of this analysis.

7.1 Combining data from different geographic areas

In Chapter 3, we developed a hidden SIRS model for cholera outbreaks, tested parameter estimation and cholera outbreak prediction using simulated data, and analyzed data from Mathbaria, Bangladesh. In Chapter 4, we laid a foundation for extending our analysis to multiple geographical areas, and in Chapter 5 we added more covariates and covariate lags to the Mathbaria analysis. Combining the ideas and framework of these chapters, next steps include an analysis combining data from all study areas in Bangladesh. Our goal is to develop a predictive model of cholera outbreaks that can be applied to other geographical areas in order to understand which covariates are universally important for prediction.

Analysis using data from just one study area is limited, especially for the thanas that only have three years of data available, as seen in Chapter 4. Therefore, developing techniques to combine data across study areas is a necessity. Development of these techniques will start with the analysis of a simple combination of data from two geographically similar

thanas under the assumption that both have the same underlying disease process, allowing for different intercepts in the environmental force of infection and accounting for different population sizes. Covariate estimates from the pooled data will be compared against estimates from the analysis of the individual thanas. Using simulated data, we will also test if and when it is possible to recover similar parameter estimates from data generated under the same parameter assumptions. We can simulate many data sets using one set of parameters for our underlying hidden SIRS model and then evaluate comparability of the parameters estimated using the separate data sets. This simulation framework will allow us to test whether the disagreement in SIRS parameter estimates across thanas, currently observed in our analyses, could be explained by the low information content of the data from each thana. Moreover, we can use these simulations to probe how informative the data need to be to recover the true underlying signal from all data sets.

The ability to combine information across study sites may improve parameter identifiability when many parameters are considered in the model, as in Chapter 5. Applying our Bayesian framework to a longer time series of cholera case counts could also allow us to estimate, rather than set, μ and to use a less informative prior on γ.

7.2 Covariate pre-processing

Another goal is to develop improved methods for combining information across sampling sites within each thana. The cubic spline method that we currently use for covariate smoothing could be causing a loss of information through over-processing the covariates. Another option might be to transform the environmental data into pseudo-covariates using a method such as principal component analysis [Jolliffe, 1986]. If we want to include many covariates from all of the water bodies, we will need to address the fact that many of the environmental measurements are capturing the same ecological process of interest. Thus, these are highly correlated, similar measurements. Including covariates from individual water bodies would lead to similar problems encountered in Chapter 5, highlighting the importance of using methods to deal with many, highly correlated predictors.

A first step at improving these methods would be to estimate the parameters in our shrinkage priors instead of using multiple different standard deviations. We could also

use priors with different shrinkage properties. Gelman et al. [2008] recommend using a Cauchy distribution as a default prior, decreasing the degrees of freedom in our t-distribution prior to 1. The horseshoe prior of Carvalho et al. [2010] performs well for sparse signals. Park and Casella [2008] explore a fully Bayesian version of the Lasso [Tibshirani, 1996], assuming conditional Laplace prior distributions on parameters. We could also shrink the effect of covariates which are less likely to be associated with cholera transmission using Zellner's g-prior [Zellner and Moulton, 1985], which assigns a normal prior distribution to the regression coefficients with variance equal to $\frac{g}{\sigma}(\mathbf{C}^T\mathbf{C})^{-1}$, where $\mathbf{C} = (\mathbf{C}_1, ..., \mathbf{C}_p)$ is a matrix of predictors, σ^2 is the sampling variability in the regression model, and g is a positive real number. Mitchell and Beauchamp [1988] use "spike and slab" priors for regression coefficients that can be deleted from the model, assigning a prior point mass at zero and a diffuse uniform prior for other values of the coefficient.

Choosing between possible predictive models will be an important step. We will develop methods for model selection to enable comparison between different models and evaluation of predictive ability to determine which set of covariates are most predictive of cholera outbreaks. Dukic et al. [2012] perform model comparison using Bayes factors, and Mukherjee and West [2009] examine posterior model probabilities. However, calculation of the marginal likelihood from the posterior is not straightforward, as discussed by Perrakis et al. [2013], so we will need to study these methods carefully. To address the need for covariate selection, reversible jump Markov chain Monte Carlo [Green, 1995] allows for movement between parameter spaces of different dimensions in MCMC samplers, and thus might be easily incorporated into our current methods.

7.3 Particle MCMC improvements

Multicollinearity of our predictive covariates could explain the poor prediction in the results of Chapter 5. The sparsity inducing priors may not be enough to fix this problem. Mixing of the Markov chain is also an issue. Recently, several papers have emerged which suggest techniques for balancing the advantages and disadvantages of increasing the number of particles in the particle MCMC [Doucet et al., 2012, Sherlock et al., 2013, Pitt et al., 2012]; implementing these techniques could improve poor mixing. Since the estimate of the

likelihood generated by the SMC algorithm converges to the true likelihood as the number of particles goes to infinity, increasing the number of particles decreases the variance in the estimated likelihood; however, this also increases computation time. The key is to strike a balance between computation time for simulations and mixing. For our current analyses, we have used 100 particles in the algorithms. Future work will incorporate some of these techniques to improve mixing while avoiding unnecessary simulation time. We will test the effects of using more particles in the SMC algorithm, especially in the analysis of the Phase 1 data where mixing is very poor.

7.4 Further modeling using a latent water compartment

In Chapter 6, we used our Bayesian approach to estimate the parameters of an SIWR model, using a latent water compartment to model the environmental contribution to cholera outbreaks. The chains are not mixing well, and large values for the concentration of *V. cholerae* in the environment, quantified by the W compartment, are contributing to prohibitively long computation times. We will explore methods for bounding the water compartment. One method would be to just put an upper bound on the size of W and test whether this would automatically rescale all the parameters without sacrificing the ability to estimate covariate effects and to perform prediction. A different time evolution of W could also be considered, such as modeling W as a continuous variable. In this hybrid approach, evolution of a continuous W would still be based on the discretely evolving SIR states. Alternatively, we could put everything on a continuous scale and use a stochastic differential equation approach, similar to Ionides et al. [2006] and Bhadra et al. [2011].

We will also test the hypothesis that the hidden SIWR model can be accurately approximated by the hidden SIRS model, as suggested by the results of Section 6.5. Thus far we have only looked at the results using one set of data simulated from the SIWR model. With more data, simulated from different sets of parameters, we will test when this approximation does not hold.

BIBLIOGRAPHY

A. S. Akanda, A. S. Jutla, and S. Islam. Dual peak cholera transmission in Bengal Delta: A hydroclimatological explanation. Geophysical Research Letters, 36(19), 2009. ISSN 1944-8007.

A. S. Akanda, A. S. Jutla, M. Alam, G. Constantin de Magny, A. K. Siddique, R. B. Sack, A. Huq, R. R. Colwell, and S. Islam. Hydroclimatic influences on seasonal and spatial cholera transmission cycles: Implications for public health intervention in the Bengal Delta. Water Resources Research, 47(3), 2011. ISSN 1944-7973.

D. F. Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. Journal of Chemical Physics, 127(21):214107, 2007.

C. Andrieu and G. O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. The Annals of Statistics, 37(2):697–725, 2009.

C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(3):269–342, 2010.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 41(1):164–171, 1970.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. Genetics, 164(3):1139–1160, 2003.

A. Bhadra, E. L. Ionides, K. Laneri, M. Pascual, M. Bouma, and R. C. Dhiman. Malaria in Northwest India: Data analysis via partially observed stochastic differential equation

models driven by Lévy noise. Journal of the American Statistical Association, 106(494): 440–451, 2011.

C. Bretó, D. He, E. L. Ionides, and A. A. King. Time series analysis via mechanistic models. Annals of Applied Statistics, 3(1):319–348, 2009.

Y. Cao, D. T. Gillespie, and L. R. Petzold. Avoiding negative populations in explicit Poisson tau-leaping. The Journal of Chemical Physics, 123(5):054104, 2005.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Biometrika, page asq017, 2010.

S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. Journal of the Royal Society Interface, 5(25):885–897, 2008.

A. Chatterjee, D. G. Vlachos, and M. A. Katsoulakis. Binomial distribution based τ-leap accelerated stochastic simulation. The Journal of Chemical Physics, 122:024112, 2005.

C. Codeço. Endemic and epidemic dynamics of cholera: the role of the aquatic reservoir. BMC Infectious Diseases, 1(1):1, 2001.

R. R. Colwell and A. Huq. Environmental reservoir of *Vibrio cholerae*: The causative agent of cholera. Annals of the New York Academy of Sciences, 740(1):44–54, 1994.

O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproductive ratio R_0 in models for infectious diseases in heterogeneous populations. Mathematical Biology, 28:365–382, 1990.

A. Doucet, N. De Freitas, and N. Gordon. Sequential Monte Carlo Methods in Practice. Springer, 2001.

A. Doucet, M. Pitt, G. Deligiannidis, and R. Kohn. Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. arXiv preprint. arXiv:1210.1871, 2012.

V. Dukic, H. F. Lopes, and N. G. Polson. Tracking epidemics with Google flu trends data and a state-space SEIR model. Journal of the American Statistical Association, 107(500): 1410–1426, 2012.

D. Eddelbuettel. Seamless R and C++ Integration with Rcpp. Springer, New York, 2013.

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. Journal of Statistical Software, 40(8):1–18, 2011.

M. C. Eisenberg, G. Kujbida, A. R. Tuite, D. N. Fisman, and J. H. Tien. Examining rainfall and cholera dynamics in Haiti using statistical and dynamic modeling approaches. Epidemics, 5(4):197–207, 2013a.

M. C. Eisenberg, S. L. Robertson, and J. H. Tien. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. Journal of Theoretical Biology, 324(0):84–102, 2013b.

B. F. Finkenstädt and B. T. Grenfell. Time series modelling of childhood diseases: a dynamical systems approach. Journal of the Royal Statistical Society: Series C (Applied Statistics), 49(2):187–205, 2000.

G. P. Garnett, S. Cousens, T. B. Hallett, R. Steketee, and N. Walker. Mathematical models in the evaluation of health programmes. Lancet, 378(9790):515–25, Aug 6 2011.

A. Gelman, A. Jakulin, M. G. Pittau, and Y. S. Su. A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2 (4):1360–1383, 2008.

M. A. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. The Journal of Physical Chemistry A, 104(9):1876–1889, 2000.

D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics, 22(4):403–434, 1976.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry, 81(25):2340–2361, 1977.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. The Journal of Chemical Physics, 115(4):1716–1733, 2001.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. Nature, 457(7232):1012–1014, 2008.

R. I. Glass, S. Becker, M. I. Huq, B. J. Stoll, M. U. Khan, M. H. Merson, J. V. Lee, and R. E. Black. Endemic cholera in rural Bangladesh, 1966–1980. Am J Epidemiol, 116(6): 959–70, Dec 1982.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732, 1995.

J. B. Harris, R. C. LaRocque, F. Qadri, E. T. Ryan, and S. B. Calderwood. Cholera. Lancet, 379(9835):2466–76, Jun 30 2012.

D. M. Hartley, J. G. Morris Jr, and D. L. Smith. Hyperinfectivity: A critical element in the ability of *V. cholerae* to cause epidemics? PLoS Medicine, 3(1):e7, 2005.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57:97–109, 1970.

D. He, E. L. Ionides, and A. A. King. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. Journal of the Royal Society Interface, 7: 271–283, June 2010.

L. Held, M. Höhle, and M. Hofmann. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Statistical Modelling, 5(3):187–199, 2005.

A. Huq, E. B. Small, P. A. West, M. I. Huq, R. Rahman, and R. R. Colwell. Ecological relationships between *Vibrio cholerae* and planktonic crustacean copepods. Applied and Environmental Microbiology, 45(1):275–283, 1983.

A. Huq, R. R. Colwell, R. Rahman, A. Ali, MA Chowdhury, S. Parveen, D. A. Sack, and E. Russek-Cohen. Detection of *Vibrio cholerae* O1 in the aquatic environment by fluorescent-monoclonal antibody and culture methods. Applied and Environmental Microbiology, 56(8):2370–2373, 1990.

A. Huq, R. B. Sack, A. Nizam, I. M. Longini, G. B. Nair, A. Ali, J. G. Morris Jr, M. N. Khan, A. K. Siddique, M. Yunus, M. J. Albert, D. A. Sack, and R. R. Colwell. Critical factors influencing the occurrence of *Vibrio cholerae* in the environment of Bangladesh. Applied and Environmental Microbiology, 71(8):4645–4654, 2005.

International Vaccine Institute. Country investment case study on cholera vaccination: Bangladesh. International Vaccine Institute, Seoul, 2012.

E. L. Ionides, C. Bretó, and A. A. King. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences, 103(49):18438–18443, 2006.

I. T. Jolliffe. Principal component analysis. Springer-Verlag, 1986.

A. Jutla, E. Whitcombe, N. Hasan, B. Haley, A. Akanda, A. Huq, M. Alam, R. B. Sack, and R. Colwell. Environmental factors influencing epidemic cholera. The American journal of tropical medicine and hygiene, 89(3):597–607, 2013.

M. J. Keeling and P. Rohani. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, 2008.

M. J. Keeling and J. V. Ross. On methods for studying stochastic disease dynamics. Journal of The Royal Society Interface, 5(19):171–181, 2008.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society of London. Series A, 115:700–721, 1927.

A. A. King, E. L. Ionides, M. Pascual, and M. J. Bouma. Inapparent infections and cholera dynamics. Nature, 454(7206):877–880, 2008.

K. Koelle and M. Pascual. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. The American Naturalist, 163(6):901–913, 2004.

K. Koelle, X. Rodó, M. Pascual, M. Yunus, and G. Mostafa. Refractory periods and climate forcing in cholera dynamics. Nature, 436(7051):696–700, 2005.

A. Kolmogorov. Uber die analytischen methoden in der wahrscheinlichkeitsrechnung. Mathematische Annalen, 104(1):415458, 1931.

I. M. Longini, M. Yunus, K. Zaman, A. K. Siddique, R. B. Sack, and A. Nizam. Epidemic and endemic cholera trends over a 33-year period in Bangladesh. Journal of Infectious Diseases, 186(2):246–251, 2002.

I. M. Longini, Jr., A. Nizam, M. Ali, M. Yunus, N. Shenvi, and J. D. Clemens. Controlling endemic cholera with oral vaccines. PLoS medicine, 4(11):e336, 2007.

R. May and R. M. Anderson. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, 1991.

P. McCullagh and J. A. Nelder. Generalized Linear Models. Chapman & Hall/CRC, 1989.

T. McKinley, A. R. Cook, and R. Deardon. Inference in epidemic models without likelihoods. The International Journal of Biostatistics, 5(1):1–40, 2009.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. Journal of Chemical Physics, 21(6): 1087–1092, 1953.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. Journal of the American Statistical Association, 83(404):1023–1032, 1988.

C. Mukherjee and M. West. Sequential monte carlo in model comparison: Example in cellular dynamics in systems biology. JSM Proceedings, Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association, pages 1274–1287, 2009.

T. Park and G. Casella. The Bayesian Lasso. Journal of the American Statistical Association, 103(482):681–686, 2008.

K. Perrakis, I. Ntzoufras, and E.G. Tsionas. On the use of marginal posteriors in marginal likelihood estimation via importance-sampling. arXiv preprint. arXiv:1311.0674, 2013.

M. K. Pitt, R. S. Silva, P. Giordani, and R. Kohn. On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. Journal of Econometrics, 171(2):134 – 151, 2012. ISSN 0304-4076.

D. A. Rasmussen, O. Ratmann, and K. Koelle. Inference for nonlinear epidemiological models using genealogies and time series. PLoS Computational Biology, 7(8):e1002136, 2011.

D. A. Sack, R. B. Sack, G. B. Nair, and A. K. Siddique. Cholera. The Lancet, 363(9404): 223–233, 2004.

R. B. Sack, A. K. Siddique, I. M. Longini, A. Nizam, M. Yunus, S. Islam, J. G. Morris, A. Ali, A. Huq, G. B. Nair, F. Qadri, Shah, M. Faruque, D. A. Sack, and R. R. Colwell. A 4-year study of the epidemiology of *Vibrio cholerae* in four rural areas of Bangladesh. Journal of Infectious Diseases, 187(1):96–101, 2003.

C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. On the efficiency of pseudo-marginal random walk Metropolis algorithms. arXiv preprint. arXiv:1309.7209, 2013.

S. Swaroop and R. Pollitzer. Cholera studies. 2. World incidence. Bull World Health Organ, 12(3):311–58, 1955.

H. M. Taylor and S. Karlin. An Introduction to Stochastic Modeling. Academic Press, 3rd edition, 1998.

T. Tian and K. Burrage. Binomial leap methods for simulating stochastic chemical kinetics. The Journal of Chemical Physics, 121:10356–10364, 2004.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

J. H. Tien and D. J. D. Earn. Multiple transmission pathways and disease dynamics in a waterborne pathogen model. Bulletin of Mathematical Biology, 72(6):1506–1533, 2010.

M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. The journal of machine learning research, 1:211–244, 2001.

WHO. Who cholera factsheet, February 2014. URL `http://www.who.int/mediacentre/factsheets/fs107/en/`.

D. J. Wilkinson. Stochastic Modelling for Systems Biology. CRC press, 2nd edition, 2011.

A. Zellner and B. R. Moulton. Bayesian regression diagnostics with applications to international consumption and income data. Journal of Econometrics, 29(1-2):187–211, 1985.